2010

Characterization of the signal sequence binding domain of Ffh by genetics and comparative analysis

Stacy Stamey Duncan
Iowa State University

Characterization of the signal sequence binding domain of Ffh by genetics and comparative analysis

by

Stacy S. Duncan

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Gregory Phillips, Co-Major Professor
Leslie Miller, Co-Major Professor
Diane Bassham
Drena Dobbs
Cathy Miller

Iowa State University

Ames, Iowa

2010

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The signal recognition particle (SRP) is a ribonucleoprotein complex whose components are highly conserved throughout all three domains of life, where it functions to target proteins to extracytoplasmic locations. In *Escherichia coli*, the SRP is comprised of a single essential protein (Ffh) in complex with a 4.5S RNA species. To better understand how the structure of Ffh contributes to its function, we have used genetic approaches to isolate and characterize new *ffh* mutants altered in two distinct domains of the protein. Both domains of interest have been implicated as being important for binding to hydrophobic signal peptides of membrane proteins. These studies include using a random sequence approach to identify amino acids important for activity of the finger loop domain. The finger loop was identified from structural analysis as a ~20 amino acid domain with the unusual properties of being both hydrophobic and exposed near the surface of Ffh. Approximately 1% of the random sequences were able to replace the FL domain of Ffh. Bioinformatic analysis of the random sequences revealed that all of the complementing sequences followed a trend of high hydrophobicity at the amino-terminus that decreased towards the carboxy end. These observations were validated by observing that mutants that deviated from this trend rendered Ffh nonfunctional. Mutants were characterized by growth rates that allowed the sequences to be grouped into three functional classes. Secondary and tertiary structure predictions suggested that the products of the random sequences lack extensive secondary structure, which is consistent with the role of the finger loop in binding a variety of ligands. To address the importance of the conserved methionine residues at the carboxy-terminal region (M-domain) in SRP function, we combined phylogenetic comparisons with functional

studies, including replacing methionine residues within the M-domain with other residues that varied in hydrophobicity, side chain flexibility and charge. These studies revealed that, in *E. coli*, the M-domain of Ffh was able to tolerate substitutions of five different hydrophobic amino acids including valine, phenylalanine, tyrosine, tryptophan and isoleucine for the conserved methionine residues found in helix αM4 and the extreme C-terminus. Phylogenetic comparisons of microorganisms with varying optimal growth temperatures revealed methionine residues were substituted with amino acids containing less flexible side chains. Interestingly, we observed that mutants containing less flexible residues were able to support cell viability at higher growth temperatures better than at lower temperatures. Phylogenetic comparisons also revealed three positions where methionine is highly conserved. We show that replacing all of the methionine residues, except these three highly conserved residues, with valine yielded a functional SRP. In contrast to predicted results, these studies reveal that the M-domain of Ffh is highly flexible in content and that methionines are not absolutely required for SRP function. Collectively, these efforts have contributed to our understanding of SRP function by identifying key features essential for the function of the signal sequence binding domain of the Ffh protein component of SRP.

CHAPTER l. General Introduction

Dissertation Organization

This dissertation is organized into six chapters including: the first chapter containing the literature review; chapters 2 and 3 which contain papers in preparation; chapter 4 containing a brief review and introduction to chapter 5 which contains a published paper, all of which cover my doctoral research; and chapter 6 which includes a general conclusion of the work and future directions.

Introduction

The following sections of this chapter will provide the reader with a review of literature, including scientific results, pertinent to the research detailed in subsequent chapters of the text. The literature review will begin with the initial discovery, using electron microscope technology, and biochemical characterization of the signal recognition particle (SRP) protein translocation pathway in eukaryotic cells. Research leading to the discovery and characterization of the bacterial SRP, including the protein component termed Ffh (fifty-four homolog), will then be discussed. Subsequent sections will focus on work elucidating the role of Ffh in SRP dependent protein localization in *Escherichia coli*, leading to our characterization of features of the Ffh protein important for its function.

Discovery of a protein translocation mechanism

Prior to the development and advancement of electron microscopy (EM) technology, details of the cellular biology of eukaryotic cells as well as bacteria remained largely unknown.

For example, eukaryotic cytoplasmic components such as ribosomes, organelles and the endoplasmic reticulum (ER) are too small to be clearly observed with conventional microscopy, but were resolved by the EM (31, 43). Also, subcellular compartments of bacterial cells, including the outer and inner membranes, remained obscure until the advent of EM.

In the early to mid-1950's George Palade used the EM to pioneer studies of protein localization. He observed in eukaryotic cells that ribosomes, typically observed in the cytoplasm, were also found attached to the membrane of the ER (43). A decade later, it was further observed that ribosomal attachment to the ER membrane was coupled with the co-translational transfer of nascent peptide chains of secretory proteins across the membrane upon exiting the ribosome (43, 50). Experimental data also revealed that synthesis of proteins retained in the cytosol occurs on free ribosomes while ribosomes synthesizing secretory proteins are recruited to the ER membrane (52). Furthermore, it was observed that nascent chains of membrane bound ribosomes were resistant to proteolysis. Specifically, the amino terminus segment of secretory proteins was resistant to digestion due to protection provided by the ER membrane. This represented the first suggestion that the amino terminus might play an important role in ribosome membrane interaction (56, 57). Blobel and Sabatini used these observations to attempt to explain why ribosomes actively translating secretory proteins are recruited to the ER membrane while ribosomes translating cytosolic proteins are not. They postulated that mRNAs destined to be translated on membrane-bound ribosomes contain a unique DNA sequence located at the 5' end which, upon emerging from the ribosome, would "signal" attachment of the ribosome to the membrane. Subsequent translocation of the protein across the ER membrane would occur co-translationally, i.e., concomitant with polypeptide elongation (57). Experimental support for their hypothesis was provided in the early 1970's.

Several research groups observed that a model secretory protein, IgG light chain from murine myelomas translated *in vitro* on free ribosomes had a higher molecular weight than the actual secreted light chain. *In vitro* synthesis of the IgG light chain on membrane-bound ribosomes, however, yielded a product with the same molecular weight as the secreted light chain product (38, 41, 58, 59, 67, 71). It was also shown that the amino terminus of the *in vitro* translation product of the light chain contained an additional ~20 amino acids not found in the secreted light chain (58). This discovery led to the proposal that the ~20 amino acid extension initiates binding of the ribosome to the ER membrane (41).

Subsequently, in 1975, to elucidate the previous findings, Blobel and Dobberstein confirmed the previously mentioned results and further demonstrated that the lower molecular weight of the secreted light chain resulted from cleavage of the additional sequence found at the amino terminus. Importantly, the processing of the amino terminus occurred only in proteins destined for secretion. They also demonstrated that processing of the additional sequence, termed the "signal sequence," occurred before completion of the nascent peptide and was coupled to protein translocation into or across the ER membrane (7). Subsequently, using a heterologous *in vitro* system comprised of translation factors from plants, ER from dogs and ribosomes from rabbits, it was shown that translocation of the protein still occurred; indicating the mRNA and not the ribosome contains the information necessary for the localization of nascent chains of secretory proteins to the ER membrane (8). In the mid 1970's Blobel and Jackson were able to show that post-translational cleavage of signal peptides of pre-secretory proteins occurred via activity of a signal peptidase (24). Ultimately, the results presented by Blobel and Dobberstein led to a more detailed version of the hypothesis that was previously proposed by Blobel and

Sabatini in 1971 and was referred to as the "signal hypothesis". This discovery ultimately resulted in Blobel being awarded the Nobel Prize in Physiology or Medicine in 1999.

Discovery and characterization of the SRP

Although the signal hypothesis was well established, very little was known about the factors involved in the co-translational transfer of nascent chains across the microsomal membrane. In an attempt to characterize these factors, Warren and Dobberstein (81) disassembled microsomal membranes into components involved in protein translocation by treating rough microsomes with high salt. Their results revealed that high salt treatment of membranes greatly reduced the protein translocation activity of the membrane, while addition of the salt extracts restored activity. They also observed that components of the salt extract were unable to re-establish protein translocation to microsomes that were treated with the protease trypsin. They proposed that proteins, not RNA, were the active components of the salt extract since the microsomes were treated with RNase prior to extraction. Based upon their results, they concluded that the extracted membrane proteins participate in signal sequence binding and binding of the ribosome nascent chain complex to the membrane and that other membrane proteins are involved in subsequent transfer of proteins across the membrane.

Using a similar approach to characterize the translocation activity of microsomes, Walter *et al*. used different concentrations of trypsin to fractionate microsomal membrane translocation activity into two components, a cytosol exposed soluble domain and a membrane-bound domain (79). They found that trypsin treatment abolished translocation activity of the membrane component; however, the activity could be restored upon addition of the soluble component. This led to their proposal that the cytosol-exposed soluble portion contains the signal sequence

and/or ribosome recognition domain and the membrane-bound component spans the membrane, which allows for protein transfer. Their conclusion conflicted with results previously reported that a peripheral not a membrane spanning protein component was involved in protein translocation (81).

In an attempt to resolve previous conflicting results, Jackson *et al*. (25) studied the components of the salt extract reported by Warren and Dobberstein and the trypsin extract reported by Walter *et al*. (79). The extracts were treated with N-ethylmaleimide, a sulphydryl modifying reagent shown to inhibit translation (29), and subsequently were unable to restore translocation activity to microsomal membranes. They further showed that untreated extracts were able to restore translocation activity to membranes that were inactivated by N-ethylmaleimide treatment. These observations led to the conclusion that both extracts contain components that are similar both structurally and functionally and contain sulphydryl on the cytoplasmic domain important for protein translocation.

Walter and Blobel (74) continued to characterize factors required for protein translocation. Using hydrophobic chromatography, they were able to purify the component of the salt extract previously found to restore translocation activity to inactive membranes (81). Using polyacrylamide gel electrophoresis, they observed that the component was comprised of six proteins thought to be found in complex with one another since they were found in stoichiometric amounts. They further observed that the complex was the only component of the salt extract found to bind the hydrophobic matrix of ω-aminopentyl-agarose, which led the authors to propose that the exposed hydrophobic region of the protein complex participates in hydrophobic signal sequence recognition (74).

In a series of three papers Walter and Blobel detailed their attempts to explicate the mechanism by which the previously purified protein complex, termed the signal recognition protein (SRP), mediates translocation of secretory proteins across the ER membrane. In the first paper, they observed that the translation of proteins was inhibited when salt extracted membranes were absent (78). To elucidate the specificity of what they observed, they added SRP complex to a translation system that lacked membrane vesicles but contained mRNAs encoding cytoplasmic proteins, α and β globin, and the secretory protein preprolactin. They observed that the SRP inhibited translation of the secretory protein but not the cytoplasmic proteins. They proposed that inhibition was due to direct interaction of the SRP complex with ribosomes translating secretory proteins. Subsequently, using radioactively labeled SRP, direct binding of the complex with ribosomes translating secretory proteins, but not cytoplasmic proteins was established. To better understand the nature of the SRP/protein interactions, they used a leucine analog, β-hydroxyl leucine, previously shown to abolish *in vitro* translocation of preprolactin when incorporated into the nascent chain (22). They found that the SRP neither bound ribosomes translating preprolactin nor inhibited translation of preprolactin containing β-hydroxyl leucine. However, once β-hydroxyl leucine was competed out with leucine, binding of the protein complex with ribosomes translating preprolactin was re-established and translation was inhibited. Additionally, it was previously shown that protein translocation was impaired when microsomal membranes were treated with N-ethymaleimide and subsequently, the impairment of the translocation mechanism was localized to SRP (74). By treating SRP with N-ethymaleimide, they found that the SRP no longer interacted with preprolactin translating ribosomes. Altogether, their results showed that the SRP binds ribosomes translating secretory proteins and suggested the signal sequence of the nascent chain mediates the binding event.

In the second paper, the role of SRP in the binding of ribosomes translating secretory proteins to microsomal membranes was investigated (77). To examine binding of ribosomes to membranes, the researchers designed an assay using an *in vitro* translation system and differential centrifugation. They indirectly measured the binding of ribosomes to the membrane by measuring the amount of mRNA in the translation system that remained once mRNA contained in membrane bound ribosomes translating secretory proteins were removed using differential centrifugation. They showed that membranes, whose translocation activity had been depleted due to salt extraction, were unable to bind to ribosomes translating preprolactin in the absence of SRP. This led to the conclusion that SRP is essential for binding of the ribosome-nascent chain complex to microsomal membranes. Furthermore, they found when β-hydroxyl leucine was incorporated into the nascent chain of a secretory protein, the ribosome-nascent chain complex no longer bound to microsomes; however this could be overcome by competing out β-hydroxyl leucine with leucine, thus validating that the nascent chain contains information (the signal sequence) required for ribosome-nascent chain complex binding to membranes.

In the third and final paper in this noteworthy series, Walter and Blobel (76) investigated their observation that SRP selectively inhibited synthesis of secretory proteins but not of cytoplasmic proteins. To examine this phenomenon, the researchers used a synchronized translation system that included a compound known to block initiation of protein synthesis and a radiolabeled methionine residue ([^{35}S] Met), allowing them to follow each step of the of protein synthesis process. They monitored the incorporation of [^{35}S] Met into preprolactin and observed a decreased rate of incorporation in the presence of SRP but without microsomal membranes, indicating arrest in translation. Subsequently, when they added salt extracted membranes, amino acid incorporation increased significantly, indicating a release in elongation arrest imposed by

the SRP. Additionally, they found that translation resumed once ribosomes bound to microsomal membranes. Further, protein synthesis was completed while the protein was simultaneously translocated into or across the membrane. Collectively, these results led to the characterization of the SRP-dependent protein localization pathway for secretory proteins that involves the following steps: first, emergence of the signal sequence from the ribosome triggers binding of the SRP to the ribosome-nascent chain complex causing an arrest in translation; second, the SRP-ribosome-nascent chain complex binds to the membrane releasing elongation arrest that allows for co-translational translocation of the protein across the membrane.

Continued characterization of the SRP led to the discovery of two additional components essential for the translocation of secretory proteins. Initially, it was thought that the SRP was comprised of only proteins since microsomes remained active even after RNase treatment (81). However, Walter and Blobel (75) discovered that a 7S RNA was indeed an essential component of the SRP complex. This led to renaming the signal recognition protein to the signal recognition *particle*. Gilmore *et al*. (19, 20) and Meyer *et al*. (40) discovered another component of the translocation pathway termed the docking protein or the SRP receptor. Both groups reported the discovery of a 72K protein (SRP receptor) located in the ER membrane required for binding of the SRP-ribosome-nascent chain complex to the membrane and the subsequent release of translation arrest and translocation of secretory proteins across the membrane.

Search for an SRP pathway in bacteria

Although the mechanism of protein translocation mediated by the SRP had been well established in eukaryotes, a similar mechanism was only later found in bacteria. Despite extensive genetic and biochemical analysis of bacterial protein export, no evidence of an SRP

resulted (7, 14, 64, 65). In contrast to eukaryotic systems, the use of genetic approaches pioneered the discovery of components of the protein export machinery in bacteria. Prior to these efforts, it was unclear how proteins were targeted to the outer membrane, periplasmic space and inner membrane. In an early study Emr *et al*. (17) constructed a gene fusion by joining the coding regions for the integral outer membrane protein LamB, the receptor for the bacteriophage λ, with the cytoplasmic enzyme β-galactosidase, encoded by *lacZ*. Since LamB was known to contain a signal sequence at the amino terminus and shown to be synthesized by membrane bound ribosomes (49), the authors reasoned that β-galactosidase would likewise be targeted outside of the cytoplasm. Indeed, when a sufficient portion of LamB was included in the hybrid protein, β-galactosidase was exported outside of the cytoplasm. Importantly, β-galactosidase no longer was functional outside of the cytoplasm, hence providing genetic selections and screens for isolation of *E. coli* mutants defective in export of the LamB-LacZ hybrid protein. Subsequent studies were instrumental in defining important features of the signal sequence for efficient export, including maintaining a sufficient level of hydrophobicity and the ability to form a α-helix (12, 39).

Further comparison of prokaryotic and eukaryotic signal sequences revealed both were similar in composition, a short hydrophilic basic region followed by a hydrophobic stretch of amino acids. Talmadge *et al*. (68) were able to show that when a eukaryotic signal sequence was fused to a periplasmic protein in *E. coli*, efficient targeting to the periplasmic space was observed, suggesting a common mechanism for protein localization.

Additional genetic approaches led to the discovery of several components of the Sec (secretion) protein export apparatus, including SecA, SecB, SecY, and SecE. However, despite attempts to the contrary, none of these proteins proved to be equivalent to the SRP (82).

Although forward genetic approaches failed to reveal the bacterial SRP, continued analysis of eukaryotic SRP components eventually lead to the discovery of homologous bacterial components shown to be involved in protein localization in *E. coli*.

Evidence of SRP in bacteria

Evidence for the bacterial SRP was first presented by homology comparisons. Using sequence similarity and structure prediction analysis it was discovered that a highly conserved region of eukaryotic 7S RNA, helix 8, was homologous to 4.5S RNA in *E. coli* (47, 66), suggesting that 4.5S RNA was a component of the protein translocation mechanism in *E. coli*. However, 4.5S RNA had previously been shown likely to be involved in protein synthesis, with no evidence for a role in protein localization (10). Subsequently, two different groups (5, 53) cloned the structural gene for SRP54 and found significant homology between this protein and two predicted gene products from *E. coli*. These included a 48K protein with unknown function designated P48 or Ffh (fifty-four homologue) and FtsY, a protein though to be involved in cell division (18). This latter protein was also strikingly similar to the α-subunit of the eukaryotic SRP docking protein (DPα or SRα). While SRP54 and Ffh shared three homologous domains, including a N-terminal domain of unknown function, a GTP-binding domain (G domain) and a methionine rich carboxy-terminal domain (M-domain), similarity to FtsY/ SRα was limited to the G domain.

While these findings were intriguing, sequence homology did not provide the experimental evidence needed to prove that an SRP protein translocation mechanism existed in bacteria. Furthermore, because genetic screens for export mutants never identified these components, the hypothesis was met with great cynicism (1, 3, 10).

Continued analysis of *E. coli*, however, eventually led to the confirmation that Ffh and 4.5S RNA comprise the SRP complex in bacteria. First, studies conducted independently by Ribes *et al*. (51) and Poritz *et al*. (46), examined whether Ffh can form a SRP-like complex in *E. coli*. Initially, they found that eukaryotic SRP 7S RNA could functionally replace 4.5S RNA in *E. coli* and that over expression of SRP54 and Ffh as well as depletion of 4.5S RNA were detrimental to cell growth, as had been observed earlier by Brown (10). Additionally, immunoprecipitation experiments demonstrated that both Ffh and SRP54 were able to bind 4.5S RNA *in vivo*. Furthermore, they found that the Ffh-4.5S RNA complex could be functionally replaced by SRP54-4.5S RNA *in vitro*. Altogether, their results demonstrated that Ffh and 4.5S RNA forms a ribonucleoprotein complex in *E. coli* and suggested it may be functionally similar to the eukaryotic SRP.

Given that SRP54 was shown to bind 4.5S RNA, Bernstein *et al*. (6) hypothesized that Ffh could bind eukaryotic SRP 7S RNA and form a chimeric SRP that would be functionally comparable to native eukaryotic SRP. Initially, to test if Ffh could bind to SRP 7S RNA, they mixed Ffh with SRP 7S RNA and five other SRP protein subunits. Using sucrose gradient sedimentation to analyze the resulting products, they found that Ffh was able to bind SRP 7S RNA as efficiently as SRP54 used in a control reaction. Additionally, using a crosslinking assay, they showed that Ffh, in place of SRP54, was able to recognize signal sequences of secretory proteins. These results provided additional evidence that Ffh and 4.5S RNA form a ribonucleoprotein complex in *E. coli* with similar function to that of eukaryotic SRP.

In 1992, studies performed by Luirink *et al*. (34) provided conclusive experimental evidence that Ffh-4.5S RNA form a functional SRP in *E. coli*. Using a photo crosslinking assay previously used to demonstrate that SRP54 binds the signal sequence of the nascent secretory

protein preprolactin (32, 33), they found that Ffh crosslinked to signal sequences of nascent secretory proteins in crude *E. coli* extracts. They also performed the crosslinking assay with free Ffh in the absence of 4.5S RNA and found that crosslinking of signal sequences with Ffh was greatly reduced. This indicated that 4.5S RNA is required for the recognition of signal sequences by Ffh. Experimental support that SRP54 and Ffh were functionally similar was further provided when the authors found that SRP54 competed with Ffh for binding the signal sequence of the preprolactin secretory protein.

Evidence that Ffh is an essential gene product in *E. coli* was first provided by Phillips and Silhavy (45). They constructed a strain of *E. coli* where the sole copy of *ffh* was placed under control of the *araB* operator and promoter and showed *E. coli* growth was dependent on the presence of arabinose. Upon depletion of Ffh by removal of arabinose they observed multiple phenotypes, including defects in cell division and a defect in signal sequence processing of several exported proteins. Their results suggested that the bacterial SRP plays a role in protein localization; however, the exact mechanism of SRP dependent protein localization required further investigation. It was also shown that genes encoding the other putative components of the SRP pathway, i.e., *ffs* and *fts*Y, are also essential for cell viability in *E. coli* (11, 35).

Role of *E. coli* SRP

One of the first studies to suggest that the SRP function in localization of a subset of *E. coli* proteins was provided by MacFarlane and Müller (37). They observed that the uptake of the lactose analog 2-nitrophenyl-β-D-galactopyronaside (NpGal) was dependent upon the proper localization and insertion of the lactose permease (LacY), a transmembrane protein, into the inner membrane. By determining the rate of NpGal uptake they were able to monitor the amount

of LacY that was being localized to the inner membrane. The uptake of NpGal was measured under two different conditions, each involving a disruption of the SRP complex. First, using a dominant lethal 4.5S RNA mutant (46) to disrupt the SRP, they observed a decrease in LacY activity. Second, by depleting cells of Ffh (45), they again observed a decrease in the levels of LacY. Using Western Blot analysis they showed that the decreases were not due to a reduction in LacY synthesis or increased protein instability. These results indicated the membrane insertion of LacY had been impaired due the inactivation of the SRP. In contrast, upon inactivating SecA function they did not observe a defect in LacY activity. Their results suggested that the *E. coli* SRP functions in targeting proteins to the inner membrane, while the Sec pathway is specific for targeting periplasmic and outer membrane proteins.

A more direct test of the role of SRP in membrane protein localization was soon made by de Gier *et al*. (15). Using the cytoplasmic membrane protein leader peptidase (Lep) as a model, they showed that localization of this protein was severely disrupted upon depletion of either 4.5S RNA or Ffh. In contrast, localization of OmpA to the outer membrane was not affected.

Several groups continued to use depletion systems to elucidate the role of each factor of the SRP pathway in targeting various proteins. Each group reported that depletion of Ffh or 4.5S or FtsY strongly affected the insertion of inner membrane proteins and only weakly affected targeting of secreted proteins (16, 60, 72, 73).

A novel genetic approach was used by Ulbrandt *et al*. (72) to elucidate the role of the *E. coli* SRP in protein targeting. This group performed a genome-wide screen in an *E. coli* SRP mutant with the prediction that when the amount of SRP is limited in the cell, overproduction of proteins whose localization is SRP dependent would titrate out the remaining SRP, hence causing a reduction in cell viability. To test this, Ulbrandt *et al*. expressed a plasmid library in

an *E. coli* strain where *ffh* was under control of the inducible *trc* (*trp-lac*) promoter and present in single copy. In screening the plasmid library for transformants that grew only in the presence of isopropyl-β-D-thiogalactopyranoside (IPTG), the authors identified eight inner membrane proteins (IMPs) that required sufficient levels of SRP for viability. The researchers referred to the resulting phenotype as "SLO" (synthetic lethality upon overexpression). It was also noted that all eight IMPs exhibiting the SLO phenotype were predicted to span the cytoplasmic membrane multiple times. Additionally, none of the proteins were predicted to contain cleavable amino-terminal signal sequences; rather their signal sequences were located in the internal membrane spanning domain, a feature common to most IMPs. This was the first evidence to suggest that SRP uses hydrophobicity to select signal sequences for protein targeting and further supported the hypothesis that the SRP is specific for localization of IMPs. Further characterization of these proteins showed that their localization to the cytoplasmic membrane required a functional SRP pathway.

An independent genetic approach was used by Park *et al*. (44) to study the role of Ffh in membrane protein targeting. By isolating a temperature-sensitive *ffh* mutant, they were able to overcome a limitation inherent to systems that used depletion of a gene product by repressing its synthesis. Park *et al*. pointed out that these systems typically require an extended growth period that may result in secondary effects that can mask the true physiological function of the protein. In the case of Ffh this is an especially relevant concern as it was shown that even low levels of Ffh can support cell viability (4) and depletion of the gene product can take multiple generations of growth. The temperature sensitive *ffh* (*ffh*^{TS}) mutant isolated by Park *et al*. resulted in the rapid inactivation of the gene product when growth was shifted from the permissive temperature of 30^{o}C to the nonpermissive temperature of 42^{o}C. Using this system, the authors clarified the

results from prior studies and showed that soon after a shift to the non-permissive growth temperature membrane protein localization was impeded, while signal sequence processing of periplasmic and outer membrane proteins remained efficient, consistent with a role of the SRP in membrane protein targeting.

Despite the insights into SRP function provided by the studies just described, a question remained as to why SRP mutants had never been isolated by previous genetic efforts to study protein localization in *E. coli*. To approach this, Tian and Beckwith revisited the use of *lacZ* gene fusions to screen for mutants with reduced efficiency of targeting a MalF-β-galactosidase hybrid protein to the inner membrane. Initially, the researchers developed a genetic screen utilizing a strain that expressed a MalF-β-galactosidase hybrid protein. They reasoned that mutations in genes necessary for targeting membrane proteins would prevent efficient insertion of the MalF-β-galactosidase protein. As a consequence, accumulation of the β-galactosidase moiety in the cytoplasm would restore a Lac^+ phenotype. However, their selection for Lac+ mutants only yielded mutations in genes involved in disulfide bond formation (70). By modifying their approach, they screened for cells that displayed only a modest increase in β-galactosidase activity. Using this approach they found mutations in genes that encode components of the *E. coli* SRP pathway, *ffh*, *ffs* and *ftsY* (69, 70). While the screen for increased enzyme activity of the MalF-β-galactosidase was ultimately a successful strategy for isolating SRP mutants, the initial studies were limited by the requirement for relatively high levels of β-galactosidase activity.

In the characterizing these new *ffh* and *ffs* mutants, Tien and Beckwith used a facile technique for monitoring membrane protein localization. Previously Jander *et al*. (27) reported that proteins tagged with the biotin-accepting domain from the 1.3S subunit of a

transcarboxylase from *Propionibacterium shermanii* (PBST), were biotinylated only when the fusion protein was localized in the cytoplasm. By creating translational fusions of PBST to the carboxy-terminus of selected membrane proteins, the location of the polypeptide could be monitored by measuring the level of biotinylation. For example, expression of the SRP-dependent inner membrane proteins FtsQ-PSBT (possessing a single transmembrane spanning domain) and AcrB-PSBT (possessing multiple membrane spanning domains) showed evidence of biotinylation in *ffh*, *ftsY* and *ffs* mutants (69, 70).

Collectively, research indicated that unlike the eukaryotic SRP, the primary function of the *E. coli* SRP is to co-translationally target inner membrane proteins. From this, it was further hypothesized that the mode of co-translational targeting as opposed to post-translational targeting is employed by the bacterial SRP in order to prevent hydrophobic segments of inner membrane proteins from aggregating or folding prematurely in the cytoplasm (48).

Signal Sequence Recognition by the SRP

Understanding details of the mechanism by which SRP recognizes hydrophobic signal peptides was approached independently by Kurzchalia *et al*. and Krieg *et al*. (32, 33), who reported a direct interaction between the signal sequence of the secretory protein preprolactin and SRP54. The researchers used an assay where they incorporated a photochemically reactive probe into two lycine residues found in the signal sequence of the nascent chain of the secretory protein preprolactin by using modified aminoacyl-tRNAs. They found that the photoreactive preprolactin signal sequence crosslinked to the 54K protein component of the SRP and also immunoprecipitated with antibodies specific for the prolactin and 54K proteins. Their results

suggested that the 54K protein of the SRP recognizes and binds to the signal sequence of secretory proteins as they emerge from the ribosome.

Inspection of the SRP/Ffh sequence led Bernstein *et al*. (5) to propose a model for signal sequence binding that suggested that the signal sequence binding region is contained in the methionine rich M-domains of these two proteins. Predictions of SRP54/Ffh secondary structure suggested the M-domain was comprised of three α-helices, two of which contain all of the conserved methionine residues of the helical region, along with other hydrophobic amino acids, clustered on one face with polar residues on the opposite face. The researchers hypothesized that the hydrophobic nature and side chain flexibility of the methionine residues along with their spatial arrangement with other hydrophobic residues form a groove on the protein surface that contributes to the plasticity of the region which would be required for binding hydrophobic signal sequences that vary in length and amino acid composition. Furthermore, they proposed that the side chains of the methionines protrude out into the groove forming what is now known as the "methionine bristle" which would allow signal sequences to bind through hydrophobic interactions.

Several independent studies revealed that the M-domain of SRP54 indeed contains the binding sites for signal sequences of secretory proteins as well as SRP 7S RNA (21, 36, 54, 84). This latter interaction was shown by using C-terminal truncated versions of SRP54, Romisch *et al*. (54) that showed that the methionine rich M-domain mediates binding of SRP54 with SRP7S RNA. They further showed that the M-domain of SRP54 is able to bind *E. coli* 4.5S RNA. In studies conducted by Zopf *et al*., (84) and High and Dobberstein (21), SRP was digested with V8 protease, which was previously shown to cleave SRP54 into two distinct fragments, the amino-terminal NG domain and the carboxy-terminal M-domain (54). Using a photo crosslinking

assay, they found that the signal sequence of preprolactin consistently crosslinked to the M-domain of SRP54. To this point, the signal sequence recognition site of Ffh had been investigated using the cleavable signal sequence of preprolactin. However, Lutcke *et al*. (36) wanted to determine if a membrane protein, IMC-CAT that contains an uncleaved signal anchor sequence also interacted with the M-domain of SRP54. Additionally, they wanted to test if the M-domain alone was sufficient for binding signal sequences. Using a similar photo crosslinking assay to that of Zopf *et al*. (84) and High and Dobberstein (21), they found that the uncleaved signal sequence of the membrane protein IMC-CAT was crosslinked to the M-domain of SRP54. Additionally, free SRP54 containing only the M-domain was found to be crosslinked to the signal sequence of both preprolactin and IMC-CAT indicating that the M-domain is sufficient for signal sequence recognition.

Structure of Ffh M-domain

In the late 90's, the crystal structure of full length Ffh from *Thermus aquaticus* was determined and provided further support for the predicted structure of the M-domain and the "methionine-bristle" hypothesis (30). The crystal structure revealed the three domains of Ffh: the N domain, the G domain and the M-domain. The M-domain was shown to contain four alpha helices, αM1, αM2, αM3 and αM4 that are arranged to form a hydrophobic groove exposed on the surface of the protein (30). As shown in Figure 1A, the proposed hydrophobic signal sequence binding groove is formed by αM1 and αM2, connected by a flexible loop called the "finger loop", and αM4. Helix αM3 contains the binding domain for the SRP RNA (80). The crystal structure along with sequence analysis showed the majority of the surface of the binding groove is comprised of highly conserved methionines as well as other hydrophobic

amino acids; however, in thermophiles such as *T. aquaticus* many of the conserved methionine residues found in mesophilic organisms are replaced by other less flexible hydrophobic residues such as leucine, isoleucine and phenylalanine. It was hypothesized that, because thermal motion is increased at higher temperatures, the need for the increased flexibility provided by the methionine side chain is eradicated (30).

Another component of the M-domain proposed to be involved in signal sequence binding is the highly conserved finger loop. Evidence for this came from the crystal structures of the Ffh M-domains from *T. aquaticus* (30) and the human SRP54 M-domain (13), as well as the structure of the complete *E. coli* SRP (2). In the crystal structure from *T. aquaticus*, Ffh was observed as a trimeric complex where two neighboring M-domains were connected such that the hydrophobic finger loop of one domain was inserted into the hydrophobic groove of the other domain. Keenan *et al*. (30) hypothesized that the observed closed conformation of the finger loop in absence of a signal sequence served to protect the hydrophobic groove from solvents. The finger loop in the crystal structure of the *E. coli* Ffh M-domain-4.5S RNA complex was shown to be a disordered region of Ffh (2). In the structure of the human SRP54 M-domain, the finger loop and α-helix 1 were shown to be inserted into the hydrophobic groove of a neighboring M-domain that was similarly hypothesized to protect the hydrophobic groove until it is displaced by a signal sequence (13). Collectively, the three structures led to the proposal that the finger loop provides additional hydrophobicity and flexibility necessary for binding a variety of signal sequences and undergoes conformational changes in the presence and absence of a signal peptide.

Continued analysis of Ffh and its interaction with other SRP components

The function of the M-domain in Ffh was further probed biochemically by Zheng and Gierasch (83) who examined the effects of 4.5S RNA binding to this region of the protein. Using a V8 protease digestion assay, they found that binding of 4.5S RNA to Ffh protected the M-domain from complete digestion by the protease and was shown to stabilize this region of the Ffh protein. Additionally, they found that 4.5S RNA binding to Ffh had no effect on signal sequence binding; however, when 4.5S RNA was absent, the M-domain was destabilized upon binding a signal sequence. Interestingly, *in vivo* studies had previously shown that Ffh was stabilized by 4.5S RNA (28).

Subsequent biochemical and structural research has continued for the past decade to focus on structural determinations of the SRP and the SRP receptor, providing new insights into the interaction of these components. Our current understanding is that SRP, bound to its cargo, forms a heterodimer with its receptor (FtsY in *E. coli*) that triggers GTPase activity of both proteins (61-63).

To begin to build a more comprehensive picture of how Ffh, 4.5S RNA and FtsY function together in SRP protein targeting, Walter and Bradshaw (9) combined reverse-genetics with biochemical analysis to study how 4.5S RNA facilitates binding of Ffh and FtsY (the SRP receptor) and whether the M-domain of Ffh is involved in Ffh-FtsY complex formation. Using site-directed mutagenesis, they mutated highly conserved regions of the Ffh protein that did not inhibit 4.5S RNA binding to Ffh. Interestingly, they found three mutations, one upstream from the finger loop domain and two in the finger loop domain, which impair SRP activity. They observed that in the presence of 4.5S RNA, the rate of complex formation between the mutant Ffh and FtsY was greatly reduced; however, they observed that binding was only slightly affected when 4.5S RNA was absent.

Very recently, the crystal structure of an Ffh-signal peptide complex was determined by Janda *et al*. (26). This structure was solved by fusing the *ffh* gene from *Sulfolobus solfataricus* to the coding sequence for the signal anchor sequence of dipeptidyl aminopeptidase B from yeast. These two components were separated by an 11-amino acid linker and the product of this chimera was designated SRP54*. The crystal structure of SRP* revealed the signal sequence binds to the hydrophobic groove formed by αM1, αM2 and αM4 and the finger loop is shown to form a "lid" above the signal peptide. Interestingly, the signal sequence was shown to extensively interact with αM4 (Figure 1B). Indeed the binding of this model signal peptide was similar to that predicted from previous structural determinations (2, 13, 23, 30, 42, 55).

The mechanisms of SRP function known to date are primarily the result of biochemical and structural analysis. To better understand how individual components of the SRP, including specific domains within the Ffh protein, contribute to SRP function, we have utilized genetic and bioinformatic approaches. The results presented in the following chapters detail these efforts and explain how they better inform us as to the function of the SRP.

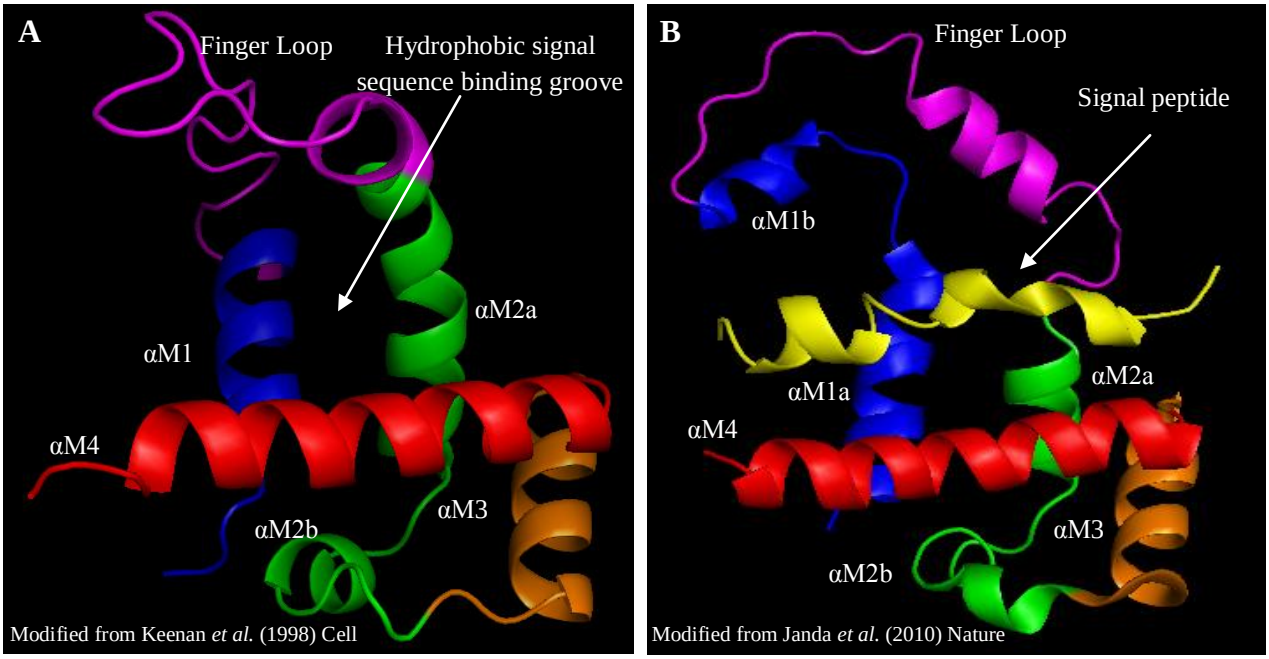

Figure 1. **Structure of the M-domain of Ffh**. **(A)** Crystal structure of the proposed hydrophobic signal sequence binding groove from *Thermus aquaticus* (Protein Data Bank 2Ffh), formed by αM1 (blue), αM2a and αM2b (green), the Finger Loop (purple), and αM4 (red). Helix αM3 (orange) contains the SRP RNA binding domain. **(B)** Crystal structure of an Ffh-signal peptide complex (Protein Data Bank 3KL4), Ffh from *Sulfolobus solfataricus* fused to the coding sequence for the signal anchor sequence of dipeptidyl aminopeptidase B from yeast. The signal sequence binding groove is formed by αM1a and αM1b (blue), αM2a and αM2b (green), the Finger Loop (purple), and αM4 (red). The signal peptide is shown in yellow.

REFERENCES

1. **Bassford, P., J. Beckwith, K. Ito, C. Kumamoto, S. Mizushima, D. Oliver, L. Randall, T. Silhavy, P. C. Tai, and B. Wickner**. 1991. The primary pathway of protein export in *E. coli*. Cell **65**:367-8.
2. **Batey, R. T., R. P. Rambo, L. Lucast, B. Rha, and J. A. Doudna**. 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science **287**:1232-9.
3. **Beckwith, J.** 1991. "Sequence-gazing?" Science **251**:1161-2.
4. **Bernstein, H. D., and J. B. Hyndman**. 2001. Physiological basis for conservation of the signal recognition particle targeting pathway in *Escherichia coli*. J. Bacteriol. **183**:2187-97.
5. **Bernstein, H. D., M. A. Poritz, K. Strub, P. J. Hoben, S. Brenner, and P. Walter**. 1989. Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature **340**:482-6.
6. **Bernstein, H. D., D. Zopf, D. M. Freymann, and P. Walter**. 1993. Functional substitution of the signal recognition particle 54-kDa subunit by its *Escherichia coli* homolog. Proc. Natl. Acad. Sci. USA **90**:5229-33.
7. **Blobel, G., and B. Dobberstein**. 1975. Transfer of proteins across membranes I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell. Biol. **67**:835-51.
8. **Blobel, G., and B. Dobberstein**. 1975. Transfer of proteins across membranes II. Reconstitution of functional rough microsomes from heterologous components. J. Cell. Biol. **67**:852-62.
9. **Bradshaw, N., and P. Walter**. 2007. The signal recognition particle (SRP) RNA links conformational changes in the SRP to protein targeting. Mol. Biol. Cell. **18**:2728-34.
10. **Brown, S**. 1989. Time of action of 4.5 S RNA in *Escherichia coli* translation. J. Mol. Biol. **209**:79-90.
11. **Brown, S., and M. J. Fournier**. 1984. The 4.5S RNA gene of *Escherichia coli* is essential for cell growth. J. Mol. Biol. **178**:533-50.
12. **Bruch, M. D., C. J. McKnight, and L. M. Gierasch**. 1989. Helix formation and stability in a signal sequence. Biochemistry **28**:8554-61.
13. **Clemons, W. M., Jr., K. Gowda, S. D. Black, C. Zwieb, and V. Ramakrishnan**. 1999. Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 Å resolution: evidence for the mechanism of signal peptide binding. J. Mol. Biol. **292**:697-705.
14. **Davis, B. D., and P. C. Tai**. 1980. The mechanism of protein secretion across membranes. Nature **283**:433-8.
15. **de Gier, J. W., P. Mansournia, Q. A. Valent, G. J. Phillips, J. Luirink, and G. von Heijne**. 1996. Assembly of a cytoplasmic membrane protein in *Escherichia coli* is dependent on the signal recognition particle. FEBS Lett. **399**:307-9.
16. **de Gier, J. W., P. A. Scotti, A. Saaf, Q. A. Valent, A. Kuhn, J. Luirink, and G. von Heijne**. 1998. Differential use of the signal recognition particle translocase targeting

pathway for inner membrane protein assembly in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **95**:14646-51.

17. **Emr, S. D., M. N. Hall, and T. J. Silhavy**. 1980. A mechanism of protein localization: the signal hypothesis and bacteria. J. Cell. Biol. **86**:701-11.
18. **Gill, D. R., and G. P. Salmond**. 1987. The *Escherichia coli* cell division proteins FtsY, FtsE and FtsX are inner membrane-associated. Mol. Gen. Genet. **210**:504-8.
19. **Gilmore, R., G. Blobel, and P. Walter**. 1982. Protein translocation across the endoplasmic reticulum. I. Detection in the microsomal membrane of a receptor for the signal recognition particle. J. Cell. Biol. **95**:463-9.
20. **Gilmore, R., P. Walter, and G. Blobel**. 1982. Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. J. Cell. Biol. **95**:470-7.
21. **High, S., and B. Dobberstein**. 1991. The signal sequence interacts with the methionine-rich domain of the 54-kD protein of signal recognition particle. J. Cell. Biol. **113**:229-33.
22. **Hortin, G., and I. Boime**. 1980. Inhibition of preprotein processing in ascites tumor lysates by incorporation of a leucine analog. Proc. Natl. Acad. Sci. USA **77**:1356-60.
23. **Ilangovan, U., S. H. Bhuiyan, C. S. Hinck, J. T. Hoyle, O. N. Pakhomova, C. Zwieb, and A. P. Hinck**. 2008. *A. fulgidus* SRP54 M-domain. J. Biomol. NMR 41:241-8.
24. **Jackson, R. C., and G. Blobel**. 1977. Post-translational cleavage of presecretory proteins with an extract of rough microsomes from dog pancreas containing signal peptidase activity. Proc. Natl. Acad. Sci. USA **74**:5598-602.
25. **Jackson, R. C., P. Walter, and G. Blobel**. 1980. Secretion requires a cytoplasmically disposed sulphydryl of the RER membrane. Nature **286**:174-6.
26. **Janda, C. Y., J. Li, C. Oubridge, H. Hernandez, C. V. Robinson, and K. Nagai**. 2010. Recognition of a signal peptide by the signal recognition particle. Nature.
27. **Jander, G., J. E. Cronan, Jr., and J. Beckwith**. 1996. Biotinylation *in vivo* as a sensitive indicator of protein secretion and membrane protein insertion. J. Bacteriol. **178**:3049-58.
28. **Jensen, C. G., and S. Pedersen**. 1994. Concentrations of 4.5S RNA and Ffh protein in *Escherichia coli*: the stability of Ffh protein is dependent on the concentration of 4.5S RNA. J. Bacteriol. **176**:7148-54.
29. **Katz, F. N., J. E. Rothman, V. R. Lingappa, G. Blobel, and H. F. Lodish**. 1977. Membrane assembly in vitro: synthesis, glycosylation, and asymmetric insertion of a transmembrane protein. Proc. Natl. Acad. Sci. USA **74**:3278-82.
30. **Keenan, R. J., D. M. Freymann, P. Walter, and R. M. Stroud**. 1998. Crystal structure of the signal sequence binding subunit of the signal recognition particle. Cell **94**:181-91.
31. **Kellenberger, E., and A. Ryter**. 1958. Cell wall and cytoplasmic membrane of *Escherichia coli*. J. Biophys. Biochem. Cytol. **4**:323-6.
32. **Krieg, U. C., P. Walter, and A. E. Johnson**. 1986. Photocrosslinking of the signal sequence of nascent preprolactin to the 54-kilodalton polypeptide of the signal recognition particle. Proc. Natl. Acad. Sci. USA **83**:8604-8.
33. **Kurzchalia, T. V., M. Wiedmann, A. S. Girshovich, E. S. Bochkareva, H. Bielka, and T. A. Rapoport**. 1986. The signal sequence of nascent preprolactin interacts with the 54K polypeptide of the signal recognition particle. Nature **320**:634-6.

34. **Luirink, J., S. High, H. Wood, A. Giner, D. Tollervey, and B. Dobberstein**. 1992. Signal-sequence recognition by an *Escherichia coli* ribonucleoprotein complex. Nature **359**:741-3.
35. **Luirink, J., C. M. ten Hagen-Jongman, C. C. van der Weijden, B. Oudega, S. High, B. Dobberstein, and R. Kusters**. 1994. An alternative protein targeting pathway in *Escherichia coli*: studies on the role of FtsY. Embo. J. **13**:2289-96.
36. **Lutcke, H., S. High, K. Romisch, A. J. Ashford, and B. Dobberstein**. 1992. The methionine-rich domain of the 54 kDa subunit of signal recognition particle is sufficient for the interaction with signal sequences. Embo. J. **11**:1543-51.
37. **Macfarlane, J., and M. Muller**. 1995. The functional integration of a polytopic membrane protein of *Escherichia coli* is dependent on the bacterial signal-recognition particle. Eur. J. Biochem. **233**:766-71.
38. **Mach, B., C. Faust, and P. Vassalli**. 1973. Purification of 14S messenger RNA of immunoglobulin light chain that codes for a possible light-chain precursor. Proc. Natl. Acad. Sci. USA **70**:451-5.
39. **McKnight, C. J., M. S. Briggs, and L. M. Gierasch**. 1989. Functional and nonfunctional LamB signal sequences can be distinguished by their biophysical properties. J. Biol. Chem. **264**:17293-7.
40. **Meyer, D. I., E. Krause, and B. Dobberstein**. 1982. Secretory protein translocation across membranes-the role of the "docking protein'. Nature **297**:647-50.
41. **Milstein, C., G. G. Brownlee, T. M. Harrison, and M. B. Mathews**. 1972. A possible precursor of immunoglobulin light chains. Nat. New. Biol. **239**:117-20.
42. **Oh, D. B., G. S. Yi, S. W. Chi, and H. Kim**. 1996. Structure of a methionine-rich segment of *Escherichia coli* Ffh protein. FEBS Lett. **395**:160-4.
43. **Palade, G. E**. 1955. A small particulate component of the cytoplasm. J. Biophys. Biochem. Cytol. **1**:59-68.
44. **Park, S. K., F. Jiang, R. E. Dalbey, and G. J. Phillips**. 2002. Functional analysis of the signal recognition particle in *Escherichia coli* by characterization of a temperature-sensitive *ffh* mutant. J. Bacteriol. **184**:2642-53.
45. **Phillips, G. J., and T. J. Silhavy**. 1992. The *E. coli ffh* gene is necessary for viability and efficient protein export. Nature **359**:744-6.
46. **Poritz, M. A., H. D. Bernstein, K. Strub, D. Zopf, H. Wilhelm, and P. Walter**. 1990. An *E. coli* ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. Science **250**:1111-7.
47. **Poritz, M. A., K. Strub, and P. Walter**. 1988. Human SRP RNA and *E. coli* 4.5S RNA contain a highly homologous structural domain. Cell **55**:4-6.
48. **Powers, T., and P. Walter**. 1997. Co-translational protein targeting catalyzed by the *Escherichia coli* signal recognition particle and its receptor. Embo. J. **16**:4880-6.
49. **Randall, L. L., S. J. Hardy, and L. G. Josefsson**. 1978. Precursors of three exported proteins in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **75**:1209-12.
50. **Redman, C. M., and D. D. Sabatini**. 1966. Vectorial discharge of peptides released by puromycin from attached ribosomes. Proc. Natl. Acad. Sci. USA **56**:608-15.
51. **Ribes, V., K. Romisch, A. Giner, B. Dobberstein, and D. Tollervey**. 1990. *E. coli* 4.5S RNA is part of a ribonucleoprotein particle that has properties related to signal recognition particle. Cell **63**:591-600.

52. **Rolleston, F. S**. 1974. Membrane-bound and free ribosomes. Sub. Cell. Biochem. **3**:91-117.
53. **Romisch, K., J. Webb, J. Herz, S. Prehn, R. Frank, M. Vingron, and B. Dobberstein**. 1989. Homology of 54K protein of signal-recognition particle, docking protein and two *E. coli* proteins with putative GTP-binding domains. Nature **340**:478-82.
54. **Romisch, K., J. Webb, K. Lingelbach, H. Gausepohl, and B. Dobberstein**. 1990. The 54-kD protein of signal recognition particle contains a methionine-rich RNA binding domain. J. Cell. Biol. **111**:1793-802.
55. **Rosendal, K. R., K. Wild, G. Montoya, and I. Sinning**. 2003. Crystal structure of the complete core of archaeal signal recognition particle and implications for interdomain communication. Proc. Natl. Acad. Sci. USA **100**:14701-6.
56. **Sabatini, D. D., and G. Blobel**. 1970. Controlled proteolysis of nascent polypeptides in rat liver cell fractions. II. Location of the polypeptides in rough microsomes. J. Cell. Biol. **45**:146-57.
57. **Sabatini, D. D., G. Blobel, Y. Nonomura, and M. R. Adelman**. 1971. Ribosome-membrane interaction: Structural aspects and functional implications. Adv. Cytopharmacol. **1**:119-29.
58. **Schechter, I**. 1975. Partial amino acid sequence of the precursor of immunoglobulin light chain programmed by messenger RNA *in vitro*. Science **188**:160-2.
59. **Schmeckpeper, B. J., S. Cory, and J. M. Adams**. 1974. Translation of immunoglobulin mRNAs in a wheat germ cell-free system. Mol. Biol. Rep. **1**:355-63.
60. **Seluanov, A., and E. Bibi**. 1997. FtsY, the prokaryotic signal recognition particle receptor homologue, is essential for biogenesis of membrane proteins. J. Biol. Chem. **272**:2053-5.
61. **Shan, S. O., R. M. Stroud, and P. Walter**. 2004. Mechanism of association and reciprocal activation of two GTPases. PLoS Biol. **2**:e320.
62. **Shan, S. O., and P. Walter**. 2003. Induced nucleotide specificity in a GTPase. Proc. Natl. Acad. Sci. USA **100**:4480-5.
63. **Shan, S. O., and P. Walter**. 2005. Molecular crosstalk between the nucleotide specificity determinant of the SRP GTPase and the SRP receptor. Biochemistry **44**:6214-22.
64. **Silhavy, T. J., P. J. Bamford, Jr., and J. R. Beckwith**. 1979. A genetic approach to the study of protein localization in *Escherichia coli*, vol. John Wiley & Sons, Inc., New York.
65. **Smith, W. P., P. C. Tai, R. C. Thompson, and B. D. Davis**. 1977. Extracellular labeling of nascent polypeptides traversing the membrane of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **74**:2830-4.
66. **Struck, J. C., H. Y. Toschka, T. Specht, and V. A. Erdmann**. 1988. Common structural features between eukaryotic 7SL RNAs, eubacterial 4.5S RNA and scRNA and archaebacterial 7S RNA. Nucleic Acids Res. **16**:7740.
67. **Swan, D., H. Aviv, and P. Leder**. 1972. Purification and properties of biologically active messenger RNA for a myeloma light chain. Proc. Natl. Acad. Sci. USA **69**:1967-71.
68. **Talmadge, K., S. Stahl, and W. Gilbert**. 1980. Eukaryotic signal sequence transports insulin antigen in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **77**:3369-73.

69. **Tian, H., and J. Beckwith**. 2002. Genetic screen yields mutations in genes encoding all known components of the *Escherichia coli* signal recognition particle pathway. J. Bacteriol. **184**:111-8.
70. **Tian, H., D. Boyd, and J. Beckwith**. 2000. A mutant hunt for defects in membrane protein assembly yields mutations affecting the bacterial signal recognition particle and Sec machinery. Proc. Natl. Acad. Sci. USA **97**:4730-5.
71. **Tonegawa, S., and I. Baldi**. 1973. Electrophoretically homogeneous myeloma light chain mRNA and its translation *in vitro*. Biochem. Biophys. Res. Commun. **51**:81-7.
72. **Ulbrandt, N. D., J. A. Newitt, and H. D. Bernstein**. 1997. The *E. coli* signal recognition particle is required for the insertion of a subset of inner membrane proteins. Cell **88**:187-96.
73. **Valent, Q. A., P. A. Scotti, S. High, J. W. de Gier, G. von Heijne, G. Lentzen, W. Wintermeyer, B. Oudega, and J. Luirink**. 1998. The *Escherichia coli* SRP and SecB targeting pathways converge at the translocon. Embo. J. **17**:2504-12.
74. **Walter, P., and G. Blobel**. 1980. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. Proc. Natl. Acad. Sci. USA **77**:7112-6.
75. **Walter, P., and G. Blobel**. 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. Nature **299**:691-8.
76. **Walter, P., and G. Blobel**. 1981. Translocation of proteins across the endoplasmic reticulum III. Signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes. J. Cell. Biol. **91**:557-61.
77. **Walter, P., and G. Blobel**. 1981. Translocation of proteins across the endoplasmic reticulum. II. Signal recognition protein (SRP) mediates the selective binding to microsomal membranes of *in-vitro*-assembled polysomes synthesizing secretory protein. J. Cell. Biol. **91**:551-6.
78. **Walter, P., I. Ibrahimi, and G. Blobel**. 1981. Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to *in-vitro*-assembled polysomes synthesizing secretory protein. J. Cell. Biol. **91**:545-50.
79. **Walter, P., R. C. Jackson, M. M. Marcus, V. R. Lingappa, and G. Blobel**. 1979. Tryptic dissection and reconstitution of translocation activity for nascent presecretory proteins across microsomal membranes. Proc. Natl. Acad. Sci. USA **76**:1795-9.
80. **Walter, P., and A. E. Johnson**. 1994. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu. Rev. Cell. Biol. **10**:87-119.
81. **Warren, G., and B. Dobberstein**. 1978. Protein transfer across microsomal membranes reassembled from separated membrane components. Nature **273**:569-71.
82. **Watanabe, M., and G. Blobel**. 1989. SecB functions as a cytosolic signal recognition factor for protein export in *E. coli*. Cell **58**:695-705.
83. **Zheng, N., and L. M. Gierasch**. 1997. Domain interactions in *E. coli* SRP: stabilization of M domain by RNA is required for effective signal sequence modulation of NG domain. Mol. Cell. **1**:79-87.
84. **Zopf, D., H. D. Bernstein, A. E. Johnson, and P. Walter**. 1990. The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be crosslinked to a signal sequence. Embo. J. **9**:4511-7.

CHAPTER 2. Essential features of the finger loop domain of Ffh as revealed by random sequences

A paper to be submitted to *The Journal of Bacteriology*

Stacy S. Duncan[1], Xiaodong Lu[2] and Gregory J. Phillips[3]

[1] Primary researcher and author, Graduate student and Professor, respectively

[2] Generated random sequence library

[3] Corresponding author and Professor

ABSTRACT

The signal recognition particle (SRP) of *Escherichia coli* is comprised of the Ffh protein in complex with 4.5S RNA. A unique feature of the Ffh protein is the "finger loop" domain, which consists of a largely hydrophobic and unstructured domain exposed near the surface of the protein. Consistent with its predicted role in binding to the hydrophobic signal sequences and transmembrane domains of SRP-dependent proteins, we show that the finger loop domain is essential for SRP function. To better understand the biological function of this domain, we developed a genetic system to screen a random sequence library to identify features of the finger loop that are important for its function. Approximately 1% of the random sequences were able to replace the finger loop domain of Ffh in *E. coli*. Bioinformatic analysis of the random sequences revealed that the single characteristic common to all of the complementing sequences was a gradient of decreasing hydrophobicity from the amino-terminus to the carboxy-terminal

end. This trend was also observed by phylogenetic comparisons of finger loop sequences from all domains of life. These observations were further validated by observing that mutations that deviated from this trend rendered Ffh nonfunctional. The random sequence mutants were also characterized by growth rates, which allowed the sequences to be grouped into three functional classes. Secondary and tertiary structure predictions suggest that the products of the random sequences lack extensive secondary structure, which is consistent with the role of the finger loop in binding a variety of ligands.

INTRODUCTION

The signal recognition particle (SRP) is a highly conserved ribonucleoprotein complex that mediates cotranslational targeting of select polypeptides to the endoplasmic reticulum in eukaryotes (52, 53) or the inner membrane in prokaryotes (11, 24, 29, 35). In eukaryotes, the SRP consists of six proteins and a single RNA (7S RNA) (52, 53), which includes a 54-kd protein that binds to the RNA and hydrophobic signal sequences (20, 26, 27, 31, 44, 57). The SRP of *Escherichia coli* is simpler, being composed of the Ffh protein (homologous to the 54-kd SRP component in eukaryotes) in complex with 4.5S RNA (42, 43). The genes encoding Ffh and 4.5S RNA (*ffs*) has been shown to be essential for cell viability (4, 41) and depletion of either of these two gene products results in a defect in localization primarily of inner membrane proteins (10, 32). Since *E. coli* expresses a "minimal" SRP it is an excellent model system to study the basic processes of membrane protein biogenesis.

The mechanism of SRP dependent targeting is similar in bacteria, eukarya and archaea. The SRP recognizes and binds to hydrophobic signal sequences of nascent peptide chains as they emerge from the ribosome. The SRP-nascent chain-ribosome complex binds to the membrane associated SRP receptor, FtsY in *E. coli* (19, 33), upon which GTPase activity of both Ffh and FtsY triggers the release of the SRP complex to participate in another round of targeting. Subsequently, the ribosome complex associates with the translocase, and the peptide chain is inserted into or across the membrane (11, 12, 14, 24).

The Ffh protein is of particular interest because it is the only SRP protein that is conserved in every organism. Both structural (2, 6, 22, 23, 25, 36, 45) and biochemical analysis (7) have revealed multiple domains of SRP54 and Ffh that function to recognize and target hydrophobic proteins to the cytoplasmic membrane. Similar to SRP54, Ffh consists of three

domains: the amino terminal N-domain, the central GTPase (G-domain), and the methionine rich carboxy terminal M-domain. The N and G domains mediate the binding of SRP with its receptor and share sequence and structural homology to similar domains in the SRP receptor, FtsY (33). The M-domain of Ffh has been shown through structural analysis and cross-linking studies to contain a binding site for 4.5S RNA as well as the primary site of signal sequence binding (31, 44, 57). The crystal structure of the M-domain from *Thermus aquaticus* revealed a deep groove formed by three alpha helices and a flexible loop that is lined with hydrophobic residues (25), a feature that was subsequently shown to be shared by SRP54/Ffh for a variety of other sources (2, 6, 22, 23, 25, 36, 45). It was recently shown that the conserved hydrophobic characteristic of the groove forms the signal sequence binding domain of the SRP (23). However, mechanistic details of signal sequence recognition and binding are not yet fully understood.

A unique feature of the binding groove of Ffh is the flexible loop region termed the "finger loop". This hydrophobic domain consists of ~20 amino acids that are exposed near the surface of the SRP and lacks significant secondary structure. The finger loop domain is part of the predicted binding site for signal sequences and hydrophobic transmembrane domains of membrane targeted proteins. Based upon crystal structures of Ffh/SRP54 from *T. aquaticus*, *E. coli*, humans and *S. solfataricus*, the finger loop adopts several different conformations suggesting this region of the M-domain modifies its conformation based upon the presence or absence of a signal sequence (2, 6, 23, 25, 45).

To better understand the role of the unique finger loop domain in SRP function we have developed a genetic system that allowed selection of sequences from a random sequence library that can replace the wild-type finger loop and restore SRP function. Random libraries have previously been used to investigate amino acid sequences important for both protein structure

and function (13, 37, 38, 48). Combined with bioinformatics analyses, we have identified heretofore undetected properties of the finger loop domain that are essential for SRP function.

MATERIALS AND METHODS

Bacterial strains, plasmids and Reagents. The strains and plasmids used in this study are shown in Table 1. All antibiotics and other chemicals were obtained from Sigma Chemical Co. (St. Louis, MO). Restriction enzymes used for cloning were obtained from New England Biolabs (Ispsich, MA) and Fermentas Life Sciences (Glen Burnie, MD). PCR primers were obtained from Integrated DNA Technologies (Coralville, IA). Antibiotics were used at the following concentrations: ampicillin (Amp), 100 µg/ml; chloramphenicol (Cam), 30 µg/ml; kanamycin (Kan), 30 µg/ml; spectinomycin (Spc), 100 µg/ml.

Plasmid construction. For these studies, we first modified pBAD*ffh*6x, a plasmid that expresses an allele of *ffh* that expresses a hexahistidine epitope tag at the carboxy terminus of Ffh (Table 1). Plasmid pBAD*ffh*6x was first modified using site directed mutagenesis to introduce a *Nhe*I restriction site at the start of the finger loop coding region, yielding pBAD*ffh*N6x (Table 1). No amino acid substitutions resulted from this change. Plasmid pBAD*ffh*N6xΔFL was made by PCR amplification using primers *ffh*NFL.S and *ffh*NFL.AS (Table 2). The PCR product was digested with *Nhe*I and *Blp*I and the gel-purified DNA was ligated into pBAD*ffh*N6x digested with the same enzymes. The resulting plasmid, pBAD*ffh*N6xΔFL, expressed Ffh such that the gene product was deleted for the finger loop encoding region (amino acids 350-370). The relevant region of each plasmid construct was confirmed by DNA sequencing (DNA Facility of Iowa State University).

Generate random sequence plasmid library. A random sequence library (18) was generated by synthesizing a randomized oligonucleotide FLrandom.S (Table 2) with an *Xba*I restriction enzyme recognition site at each end. The DNA was converted to double-stranded by first annealing the primer *ffh*FL2.S (Table 2) and extending the duplex with Klenow enzyme in the presence of dNTPs. The resulting DNA molecules were purified using Qiaex II elution kit (Qiagen, Valencia, CA) and digesting with *Xba*I. After agarose gel purification the digested DNA was ligated to pBAD*ffh*FL6x (Table 1) that had been digested with *Nhe*I and treated with antarctic phosphatase (New England Biolabs, Ispisch, MA). Ligation reactions were transformed into DH5α and plasmid DNA was isolated from pooled transformants.

Complementation tests. SLD106 was constructed by transforming PMI105 (Table 1) with plasmid p*ffh*TS-Spc, a derivative of a temperature-sensitive cloning vector based on the pSC101 origin of replication (40), which confers Spc^R. Antibiotic transformants were selected at 30^{o}C and screened for loss of Cam^R, indication loss of pFfhTS29, which also expresses a functional copy of *ffh* (Table 1).

The random library was transformed into SLD106, as described by Peterson and Phillips (39). Transformants were plated in duplicate on LB agar plates supplemented with Amp and incubated at 30^{o}C and 42^{o}C. Transformants were restreaked on LB agar plates supplemented with Spc and incubated at 30^{o}C to test for Spc^S, indicating loss of the temperature-sensitive plasmid. Plasmid DNA was isolated using a Qiagen miniprep kit (Qiagen, Valencia, CA) and the DNA sequences of the *ffh* alleles determined.

Sequence analysis. The multiple alignment tool Clustal W (49) was used to align the finger loop sequences of 109 Ffh/SRP54 proteins representing multiple species (1), as well as the predicted amino acid sequences of each mutant. ExPASy (Expert Protein Analysis System)

proteomics and sequence analysis tool, ProtScale (17), was used to determine the hydrophobicity across the finger loop sequences. ProtScale parameters used to generate hydrophobicity plots included using the Kyte and Doolittle scale and a sliding window of size 5. Additionally, hydrophobicity plots were re-generated and compared using Microsoft Office Excel 2007. Predict Protein (Structure Prediction and Sequence Analysis Service) (54) was used to generate secondary structure predictions. The homology modeling server CphModel Server 2.0 (30) was used to generate 3D predictions of the mutant sequences and each predicted model was viewed using the molecular visualization software RasMol.

Construction of additional finger loop mutants. Four additional finger loop mutants were constructed using PCR (LM-AA, LM-SS) or by replacing relevant regions of the finger loop sequence with oligonucleotides (PG-AA, DNK-LQL). To construct the LM-AA and LM-SS alleles, pBAD*ffh*N6x as a template for PCR using primers *ffh*FL-LM-AA.S, *ffh*FLAS.AS and *ffh*FL-LM-SS.S, *ffh*FLAS.AS, respectively (Table 2). Gel purified PCR products were digested with *Nhe*I and *Sac*I and ligated into pBAD*ffh*N6x digested with the same enzymes.

To construct the PG-AA, two complementary oligonucleotides (Table 2) were synthesized and mixed at a 1:1 molar ratio. The DNA was heated to 75^{o}C followed by slow cooling to room temperature resulting in a double stranded molecule with 4-base overhangs compatible with the *Nhe*I restriction site. The DNA was ligated into pBAD*ffh*N6x that had been digested with *Nhe*I and treated with antarctic alkaline phosphatase (Fermentas Life Sciences, Glen Burnie, MD). Ligation reactions were digested with *Nhe*I to further reduce the background of recircularized plasmids. A similar strategy was used to construct the DNK-LQL allele, using oligonucleotides *ffh*FL-DNK-LQL.S and *ffh*FL-DNK-LQL.AS (Table 2). All recombinant plasmids were confirmed by DNA sequencing.

Growth characterization of finger loop mutants. To characterize growth of the finger loop mutants, SLD108 was constructed. This strain is deleted for genes whose products are necessary for arabinose transport (*araFGH, araE*) and utilization (*araBAD*). In addition, SLD108 expresses a mutant LacY permease that allows homogenous uptake of arabinose throughout the population (3, 34). As a result, arabinose acts as a true gratuitous inducer and the heterogeneity of gene expression of genes under *araC* control is eliminated in SDL108 (34). To construct SLD108, ECF529 (3) was first modified by lambda Red homologous recombination to inactivate *bla* (Amp^R), encoded on the chromosome, by replacement with a Kan^R gene cassette amplified from pKD4 (9). The gene cassette was subsequently deleted using Flp-mediated site-specific recombination (9). The resulting Kan^S strain, XLU102 (Table 1) was subsequently transformed with p*ffh*TS-Spc, and the *ffh*::*kan*1 allele introduced by P1 transduction (39). Like SLD106, SLD108 is inviable at 42^{o}C since the sole functional copy of *ffh* is expressed from a temperature-sensitive plasmid.

Plasmids expressing functional finger loop mutants and control plasmids were transformed into SLD108 and Amp^R colonies restreaked on LB+Amp agar plates and incubated at 42^{o}C. Single colonies were used to inoculate 5 ml of LB agar and grown at 42^{o}C for 12 h. 500 µl of the overnight cultures were transferred to 50 ml of LB medium and grown at 42^{o}C to an OD_{600} 0.5-0.6. Cultures were diluted to 10^{-8} and spotted onto 3.5 x 3.5-inch square plates containing LB+Amp media and supplemented with L-arabinose at concentrations of: 0.0002%, 0.002%, 0.004%, 0.006%, 0.008% and 0.02%. Plates were incubated at 30^{o}C for 12 h.

Bioscreen C assay to classify sequences. The kinetics of growth of the mutants was determined using a Bioscreen C Automated Growth Curve Analysis System (Growth Curves USA, Piscataway, NJ). Mutants were transformed into SLD108 and transformants were restreaked on

LB+Amp agar plates and incubated at 42^{o}C for 12 h. Single colonies from each mutant was inoculated into LB+Amp broth (1 ml), loaded into individual wells in a 96 well plate and incubated in a 42^{o}C water bath with shaking for 12 h. Overnight samples were diluted 1:l00, added to 250 µL of LB+Amp medium and loaded in triplicate into bioscreen micro-plates (10x10 wells). Bioscreen C conditions were programmed to maintain a constant growth temperature of 42^{o}C and to record the OD_{600} of each sample every 15 min over 24 h. Subsequently, growth curves were plotted and the growth rate constants were determined and analyzed using Microsoft Excel.

Detection of Ffh mutant proteins. Plasmids pBAD*ffh*N6xFL-PG-AA, pBAD*ffh*N6xFL-LM-AA, pBAD*ffh*N6xFL-LM-SS, pBAD*ffh*N6xFL-DNK-LQL along with the wild type control pBAD*ffh*N6x were transformed into SLD106. Cells were grown at 30^{o}C in LB+Amp to mid-logarithmic phase and subsequently induced with arabinose. Upon reaching an OD_{600} of ~0.6, 500 µl of culture was pelleted at 13,000 rpm in a microcentrifuge. The pellets were resuspended in 100 µl of nanopure water. 40 µl of the suspension was mixed with 38 µl of Laemmli sample buffer (Bio Rad, Hercules, CA) and 2 µl of β-mercaptoethanol. Samples were heated for 5 minutes at 95^{o}C, and then recentrifuged before loading a portion of the supernatant on two 12% polyacrylamide gels. Proteins were detected on two different gels using either the Bio-Safe Coomassie G-250 Stain (BioRad, Hercules, CA) or InVision His-Tag In Gel Stain (Invitrogen, Carlsbad, CA), which was used to detect the a hexahistidine epitope tag at the carboxy terminus of Ffh.

RESULTS

The finger loop domain is essential for Ffh function. To characterize the finger loop domain of Ffh, we first constructed SLD106 to facilitate performing complementation tests (Materials and Methods). As shown in Fig. 1A, this conditional strain grows at 42^{o}C only if it is transformed with a plasmid expressing a functional copy of *ffh*. For this, we used pBAD-*ffh*N6x, a ColE1-derivative plasmid that expresses *ffh* under control of the *araBAD* operator and promoter (Table 1). Ffh is expressed from this plasmid with a hexahistidine epitope tag on the C-terminus which was shown previously not to interfere with SRP function (39). To determine the importance of the finger loop domain for Ffh function, we used site-directed mutagenesis to modify pBAD*ffh*N6x by deleting a 60-bp region that encodes the finger loop (56), yielding pBAD*ffh*N6xΔFL (Table 1 and Fig. 1A).

In characterizing the wild type control (expressed from pBAD*ffh*N6x, Table 1), we observed, conveniently, that arabinose was not needed to complement *ffh*::*kan*1 in SLD106 when grown at 42^{o}C, apparently due to leaky expression from the *araBAD* promoter at the elevated growth temperature. Previous studies (43), as well as our own observations, revealed that elevated expression of Ffh is detrimental to cell growth, i.e., Ffh over expression exerts a dominant negative phenotype. Similar to wild type, *E. coli* expressing increased levels of FfhΔFL from pBAD*ffh*N6xΔFL grew poorly, indicating that other features of the protein are important for the dominant negative phenotype (data not shown). In contrast to the wild type control, however, expression of *ffh*ΔFL failed to complement *ffh*::*kan*1 in SLD106 at 42^{o}C (Fig. 2A and 2B).

Sequence analysis of FL from multiple species. The finger loop domain is unstructured in SRP crystallographic studies (11), and most of the amino acids within this region are not highly

conserved. To determine the extent of sequence conservation in the finger loop, we compared amino acid sequences from this domain from 109 distinct species, representing all 3 domains of life using a multiple sequence alignment (49). We found two amino acids, Pro-355 and Gly-356 from *E. coli*, were highly conserved (data not shown).

Selection of random sequences that restore finger loop function. Based upon the results of our sequence comparisons across multiple species, we chose to select random sequences that could restore function to the *ffh*ΔFL mutant as a means to investigate the features of the finger loop domain important for Ffh activity. As summarized in Fig. 1B, we generated a library of *ffh* variants where the 60-bp region encoding the finger loop was replaced by randomized bases. To isolate functional *ffh* mutants, we introduced the random sequence library into SLD106 and selected transformants that grew at 42^{o}C (Fig. 1).

To eliminate false positives, we tested each transformant that grew at 42^{o}C for Spc sensitivity, indicating loss of p*ffh*TS-Spc. Each Spc^{S} transformant was further tested to confirm its ability to grow at 42^{o}C (Fig. 2B). By comparing the efficiency of transformation at 30^{o}C and 42^{o}C we estimated that ~1% of the total number of random sequence inserts were capable of restoring function to the *ffh*ΔFL mutant. Of these, 42 complementing clones were selected for further characterization.

Growth assays to classify functional finger loop sequences. In other studies where *ffh* mutants were characterized, we observed a correlation between SRP function and growth rates (Duncan and Phillips, manuscript in preparation). While performing complementation tests of the finger loop mutants we observed differences in colony sizes among the complementing clones. To quantify the differences in growth rate reflected by colony sizes, we used a Bioscreen C Automated Growth Curve Analysis System to determine growth rate constants. From these

results we were able to divide each of the complementing finger loop clones into three distinct groups based on their growth rates (Fig. 3).

As an independent means to measure bacterial growth, we also took advantage of the ability to regulate expression of *ffh* with arabinose to determine the amount of inducer needed for each finger loop clone to complement *ffh*::*kan1*. For this, SLD108 was constructed by using a genetic background that allows homogeneous cell-to-cell expression in the presence of different arabinose concentrations (34). SLD108 was transformed with each finger loop clone and grown at 42^{o}C to cure the cells of p*ffh*TS-Spc. Spc^{S} transformants were then cultured at 30^{o}C on LB+Amp plates in the presence of different concentrations of arabinose. We observed a general trend that finger loop mutants with smaller colony sizes and slower growth rates required a higher concentration of arabinose to support growth (Fig. 3). As explained above, arabinose concentrations above 0.001% were detrimental to cell growth. Taken together, these results indicate sequences that function poorly as finger loop replacements required higher levels of expression than sequences that function at near wild type levels (Fig. 3).

Sequence analysis of finger loop clones. The DNA sequences of 42 complementing mutants were determined. The predicted amino acid sequences encoded by each random sequence are shown in Table 3. Each amino acid sequence was examined for several properties for comparison with the wild type finger loop sequence from *E. coli*.

Initially, finger loop sequences were analyzed using multiple alignment tools, including ClustalW (49), to compare the complementing mutants. We failed to observe a correlation of amino acid conservation or distribution at any of the residue positions between wild type and mutant finger loop sequences (Table 3). Since the finger loop is unusual in that it is an exposed hydrophobic domain (1), we analyzed the random sequences using hydrophobicity plots to

determine the distribution of hydrophobic amino acids across the finger loop from amino (N)-terminus to carboxy (C)-terminus. From these plots we observed that, like the *E. coli* finger loop, all of the complementing sequences consistently progressed from high hydrophobicity at the N-terminus to significantly lower hydrophobicity towards the C-terminus (Fig. 4A). Although the data in Fig. 4A utilized the Kyte-Doolittle scale (28), this trend held when other hydrophobicity scales where also tested (15, 16, 21) (Materials and Methods). In addition, although the overall trend across the finger loop was toward less hydrophobicity, a periodicity alternating between hydrophobic and hydrophilic values was also observed in the wild type finger loop sequence from *E. coli*; however, the periodicity was not as apparent in the sequences of the functional clones when analyzed collectively using the average hydrophobicity.

To determine if the trend from more to less hydrophobicity, as well as the periodicity of the residues, was conserved across phylogenetic lines, we re-examined the multiple alignment previously generated to compare the finger loop domain from 109 Ffh sequences from multiple species, representing all three domains of life (1). We used ExPASy sequence analysis tool ProtScale (17) to generate hydrophobicity plots to observe the distribution of hydrophobic amino acids within the finger loop from multiple species. Similar to the functional random sequences, we observed the progression of more to less hydrophobicity from N- to C-terminus is phylogenetically conserved, as was the alteration between hydrophobic and hydrophilic amino acids observed in the wild type finger loop sequence from *E. coli* (Fig. 5). Furthermore, our sequence analysis revealed that there is significant sequence variation in the finger loop domain from multiple species toward the C-terminus; however, the physicochemical properties of the amino acids are similar. Using Jalview 2.4.0 (5, 55) and WebLogo (8, 47), a consensus sequence was determined based upon the conservation of these properties (Fig. 5 and Table 3).

To determine what features encoded by the random sequences are most important for optimal function of Ffh, hydrophobicity plots were generated for each of the three groups of functional finger loop mutants (Fig. 4B). Although the trend from high to low hydrophobicity was conserved among the groups, there was a marked difference in the average hydrophobicity of each group at several amino acid positions (Fig. 4B). Specifically, the average hydrophobicity of the residues at several positions (2, 6, 7, 9, 10, 13, 15, and 17) were more similar between group I sequences and the wild type sequence. Group II sequences were more similar to wild type than sequences from group III at positions 1, 11 and 17. Group I sequences also more prominently revealed the periodicity observed with the wild type finger loop sequence of alternating between hydrophobic and hydrophilic residues. These observations reveal that the hydrophobicity of each amino acid position throughout the finger loop is more important for Ffh function than specific amino acid identity.

Sequence analysis revealed that the finger loop from multiple species contained highly conserved Pro and Gly residues (Table 3). Although the majority of the complementing finger loop sequences possess either a single or multiple Pro or Gly residues, they are not found at consistently conserved positions. Additionally, not all of the functional random sequences require Pro or Gly residues. For example, finger loop sequences II-12-5, II-12-7 and II-701-7 lack a Pro or Gly residue; both are members of group II by complementation tests (Table 3).

Additional sequence analysis. We further analyzed the mutant finger loop sequences to detect potential secondary structure by using PredictProtein (46). Our results reveal that all functional sequences lacked significant secondary structure, consistent with structural and biochemical analysis of Ffh (2, 6, 7, 22, 23, 25, 36, 45). Subsequently, we used DisEMBL and DISOPRED2 (54) to predict if the finger loop mutants, like wild type, are disordered in the Ffh protein.

Results revealed that each of the mutants has an increased probability to be a disordered region in Ffh (data not shown).

From the crystal structure of the Ffh proteins from *Thermus aquaticus* (25) and *Sulfolobus solfataricus* (23), a short alpha-helical region in the finger loop domain was detected. To test if any of the sequences from the complementing mutants were predicted to form a similar structure or additional structure, we used CphModel Server 2.0 (30) to predict the 3-D structure from both the complementing mutants. Each of the functional finger loop mutants was predicted to contain the short alpha-helical region detected in *T. aquaticus* and *S. solfataricus*, however, no additional structure was predicted.

Functional characterization of the finger loop domain. The importance of multiple features of the finger loop revealed from analysis of the random sequences was directly tested. First, to test the importance of hydrophobic amino acids at the extreme N-terminus, we used site-directed mutagenesis, as described in Materials and Methods, to construct mutants where the first two hydrophobic residues were replaced with either alanine (L350A and M351A) or serine (L350S and M351S). As shown in Fig. 2C and 2D, both mutations rendered *ffh* non-functional. Next, the contribution made by the "gradient" of hydrophobicity through the finger loop on Ffh function was tested by replacing three C-terminal amino acids with hydrophobic residues (D362L, N363Q, K365L). This mutant also failed to complement *ffh::kan1* (Fig. 2C and 2D).

Site directed mutagenesis was next used to construct additional mutants where the conserved proline and glycine were replaced with alanine (P355A and G356A) to assess the importance these amino acids in Ffh function. As shown in Fig. 2D, this mutant again rendered Ffh non-functional. To confirm that Ffh was being synthesized in each of the mutants we took

advantage of a hexahistidine epitope tag present at the C-terminus of the full length Ffh protein to visualize each polypeptide by SDS-PAGE (Fig. 6).

Although the hexahistidine tag was shown not to interfere with Ffh function when expressed from pBAD*ffh*N6x, we also tested each of the finger loop mutants when expressed without the epitope tag. For these studies we constructed plasmid derivatives to express the Ffh mutants without the hexahistidine epitope tag. Even when expressed without the epitope tag, none of the mutants tested complemented in SLD106 (data not shown). However, elevated expression of each of the specific finger loop mutants at the highest concentrations of arabinose significantly inhibited growth, indicating that all expressed mutant Ffh protein.

DISCUSSION

The crystal structure of the Ffh M-domain from *Thermus aquaticus* (25) as well as the full length Ffh proteins from other species (2, 6, 22, 23, 36, 45) consistently revealed the M-domain consists of anywhere from four (in *T. aquaticus*) to seven (in Human SRP) α-helices with the predicted signal sequence binding groove formed by several of these helices in combination with the flexible finger loop domain. Most recently, the structure of a signal peptide bound to this hydrophobic groove was presented, confirming the importance of this feature of the protein (23). In this structure of *Sulfolobus solfataricus*, the finger loop appeared to participate in signal peptide binding by forming a "lid" on top of the signal sequence binding domain (23). Consistent with this important role in signal sequence binding, we showed that the finger loop is essential for Ffh function, hence viability of *E. coli* (Fig. 2A and 2B).

Initially, to identify amino acid residues that might be conserved in the finger loop domain, we used using multiple sequence alignments to compare 109 finger loop sequences from

multiple species. From this, two residues, (Pro-355 and Gly-356 from *E. coli*) were found to be highly conserved in the finger loop; however, most of the amino acid positions throughout the finger loop domain were variable.

As an independent approach to identify specific features of the finger loop essential for Ffh function we selected random sequences that were able to functionally replace the finger loop. The use of random sequence libraries has been used for various applications such as determining residues that contribute to protein or enzyme function and structure as well as to select for new mutants that vary in biological activity, such as binding alternate proteins and receptors (13, 37, 38, 48). Characterization of random sequences with biological function have effectively been used to identify characteristics of signal sequences found in secretory proteins (37), as well as to identify important amino acids for DNA Pol I (38, 48). To apply this approach to characterize Ffh, we screened a random sequence library generated by randomizing 20 residues in the finger loop domain and selecting clones capable of complementing the *ffh* deletion mutant in SLD106 (Fig. 1). Our results revealed that a surprisingly high number (~ 1%) of randomly generated sequences were able to replace the finger loop domain of Ffh, of which 42 independent clones were selected for further analysis.

Although all of the selected clones encoded functional Ffh protein, we observed differences in growth, as initially measured by colony sizes, indicating the mutants functioned at different levels of efficiency. As a result, growth differences were quantified by determining growth rate constants for each mutant. From this data, we partitioned the complementing clones into three groups (Table 3). To further characterize these mutants, we reasoned that mutants displaying a slower growth rate should require increased concentrations of arabinose since expression of the alleles was regulated by the *araBAD* operator and promoter. As shown in

Table 3, our prediction was confirmed as we found that, in most cases, mutants displaying slow growth rates needed increased amounts of arabinose to support cell viability (Fig. 3).

To identify amino acids that may have been conserved at specific positions along the finger loop sequences of our complementing mutants, multiple sequence alignments were performed; however, no specific amino acids were consistently found at any of the positions (Table 3). Additionally, although Pro and Gly residues are highly conserved in the finger loop domain in nature, this conservation was not held among the random sequences (Table 3).

Since the finger loop is largely hydrophobic (2, 6, 7, 22, 23, 25, 36, 45), we next generated hydrophobicity plots using ExPASy sequence analysis tool ProtScale (17) using several different hydrophobicity scales (15, 16, 21, 28) to examine the frequency and distribution of hydrophobic amino acids in the mutants. This approach revealed a trend of increased hydrophobicity at the N-terminus and a decrease in hydrophobicity at the C-terminus in the finger loop mutants, as well as in the wild type sequence from *E. coli* (Fig. 4A). Furthermore, we observed in the wild type sequence and in 80% of the sequences of the complementing clones, a periodic distribution of hydrophobic amino acids. As shown in Fig. 5, the trend toward less hydrophobicity was also shared among the three groups of functional finger loop sequences and the periodicity of alternating hydrophobic/hydrophilic amino acids was observed, albeit to a lesser extent, in the amino acid sequences that comprised group I. These results suggest that although the periodicity may be important for optimal function of the finger loop, it is not an absolute requirement.

The importance of the trend from high to low hydrophobicity was further supported by observing the same relationship among Ffh sequences from multiple species (Fig. 5). Hydrophobicity plots revealed phylogenetic conservation of the hydrophobic trend observed in

the functional finger loop random sequence mutants as well as the periodicity of the hydrophobic residues found in the wild type finger loop sequence (Fig. 5). Additionally, although there is little sequence homology among the finger loop sequences from complementing clones as well as different species, we observed a similarity in the physicochemical properties of the residues among sequences from different species, as shown in the consensus sequence (Table 3 and Fig. 5). Here, we see that the N-terminus contains conserved hydrophobic amino acids (positions 1, 2, 5 and 8), residues whose properties vary (positions 11-15) and the C-terminus is comprised of polar amino acids (positions 16-20). Differences in several amino acid positions among the three groups of finger loop mutants (Fig. 4B) were observed when we compared the average hydrophobicity of each group. We found that although the identity of the residues themselves varied, the hydrophobicity of amino acids found at key positions in the mutant sequences correlated with those found in the corresponding positions in the wild type finger loop. These results indicate that the chemical properties of the amino acids at selected positions, rather than the identity of the amino acids themselves, is the key determinant for function of the finger loop domain. This feature is consistent with a role of the finger loop in recognition of hydrophobic signal sequence as it is well established that signal sequences also vary in amino acid composition while retaining a hydrophobic nature (50, 51, 56). We further note that the sequences represented by the group I mutants more closely resembled wild type with respect to the overall trend in reduced hydrophobicity from N- to C- terminus, as well as the hydrophobic/hydrophilic periodicity of the residues (Fig. 4B).

To test the importance of hydrophobicity in finger loop function, we constructed a series of mutants that altered specific features in this domain predicted to be important based on the random sequence analysis. This included mutants with reduced hydrophobicity at the N-

terminus (L350A, M351A) and (L350S, M351S), increased hydrophobicity toward the C-terminus (D362L, N363Q, K365L). None of the mutants supported cell viability, confirming the importance of the hydrophobic trend discovered with the complementing random mutants. In addition we altered the conserved Pro and Gly residues (P355A and G356A) to test the importance of these amino acids. This mutant also failed to support Ffh function. With this in mind, we determined the distribution of each of these residues in the mutant finger loop sequences. We found that, while the majority of the random sequences included Pro and Gly at various positions throughout the sequence, some sequences lacked these residues. This suggests that Pro and Gly are not required for Ffh function when found in combination with other residues.

Consistent with structural and biochemical analysis of Ffh that revealed the finger loop domain is unstructured (2, 6, 7, 22, 23, 25, 36, 45), the functional finger loop mutants were also predicted to lack secondary structure and disordered in the Ffh protein (46, 54). Although the finger loop was predicted to be largely unstructured, the crystal structure of Ffh from *Thermus aquaticus* (25) and *Sulfolobus solfataricus* (23) revealed a short α-helix contained at the beginning and the center of the finger loop domain. Based upon the predicted 3-D structure of each mutant (30), a difference between the predicted structure of our functional finger loop mutants and the crystal structure from *T. aquaticus*, was not observed, indicating this is not likely an essential feature of the finger loop in all sources of Ffh.

To detect additional patterns in the finger loop sequences from complementing clones that could reveal important features of the domain, we examined the distribution of amino acids based upon size, charge, polarity and propensities toward secondary structure. No additional properties common to all of our functional finger loop sequences were observed, however.

In conclusion, this study provides another example of how screening of random sequence libraries can yield new insights into biological function. This approach has revealed that while the finger loop is extremely tolerant of amino acid substitutions, it maintains a strict requirement for hydrophobicity only at the N-terminal region of the domain. Future studies will include biochemical analysis of the products of the mutants to more precisely identify how changes to the finger loop affect SRP function.

ACKNOWLEDGEMENTS

Funding for this research was from the National Institutes of Health grant R01 GM069628.

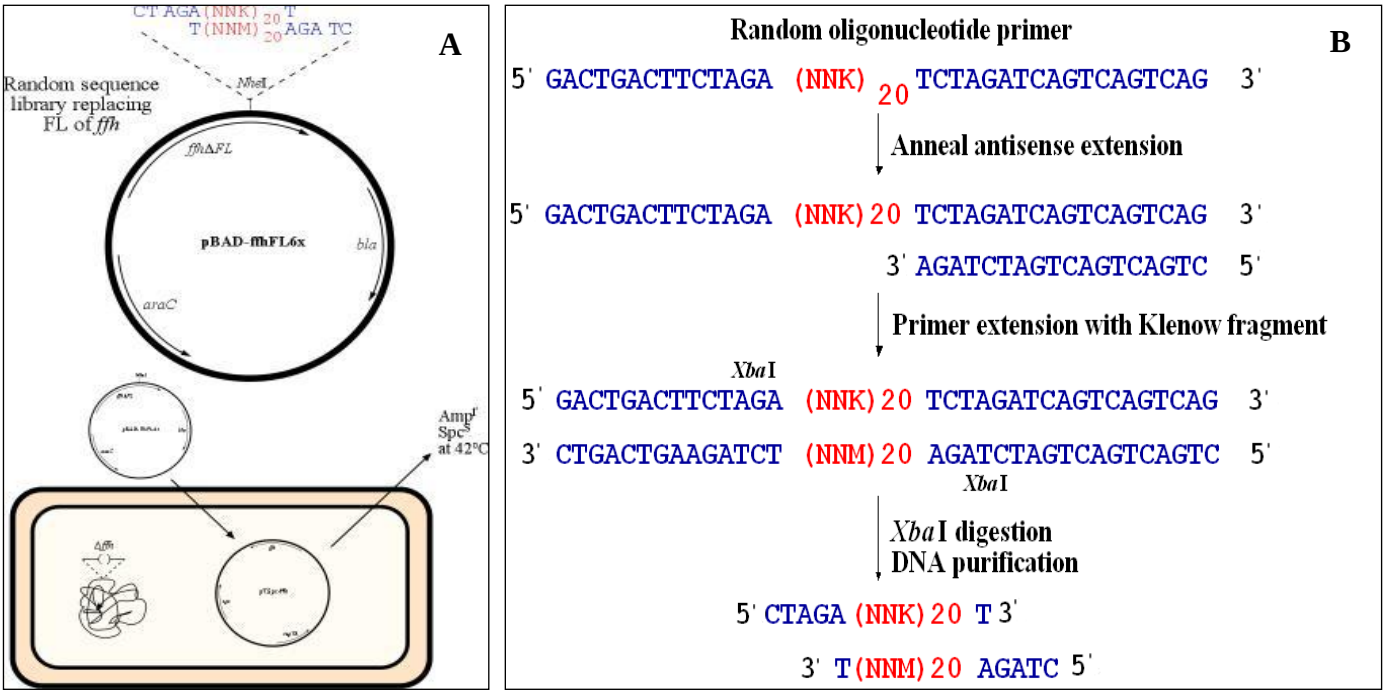

Fig. 1. (A) Genetic system for screening a random sequence library for complementing clones. The random library was transformed into SLD106 by selecting Amp^R at 42°C. Transformants were re-tested for their ability to complement the *ffh* deletion in SLD106, and for Spc^S, indicating loss of the temperature-sensitive plasmid in this strain. **(B)** Synthesis of random sequences for library construction. An oligonucleotide containing randomized codons was synthesized and converted to a double-stranded molecule, as described in Materials and Methods. The DNA was digested with *Xba*I and ligated to pBAD-*ffh*N6X at a unique *Nhe*I site engineered in place of the finger loop-encoding region.

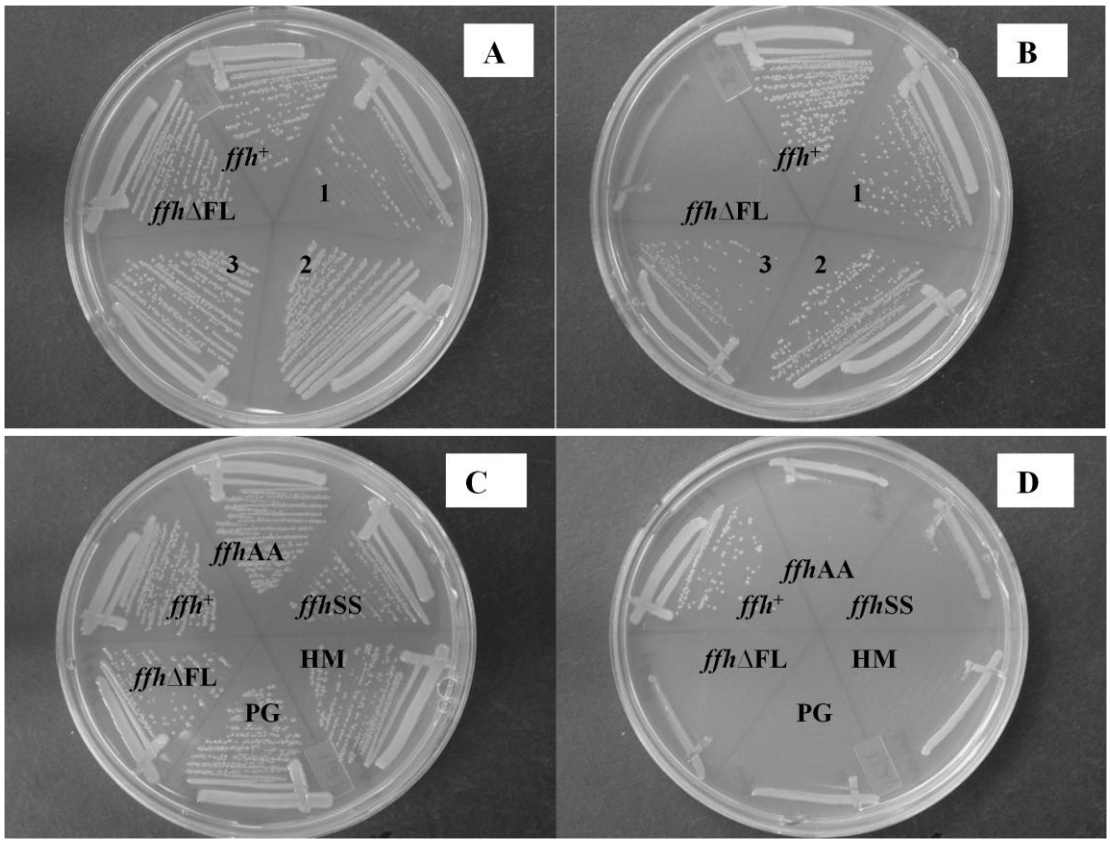

Fig. 2. Phenotypes of *ffh* finger loop mutants. (**A)** Colonies grown at 30^{o}C showing positive (*ffh*$^{+}$) and negative (*ffh*ΔFL) controls and examples of 3 mutants (1, 2 and 3). (B) Colonies restreaked at 42^{o}C from (A) showing complementation by mutants 1, 2 and 3, and *ffh*$^{+}$. (C**)** Colonies grown 30^{o}C showing positive (*ffh*$^{+}$) and negative (*ffh*ΔFL) controls and mutants *ffh*AA (L350A and M351A), *ffh*SS (L350S and M351S), PG (P355A and G356A) and HM (D362L, N363Q, K365L). (D) Colonies restreaked at 42^{o}C from (C) showing complementation by only *ffh*$^{+}$.

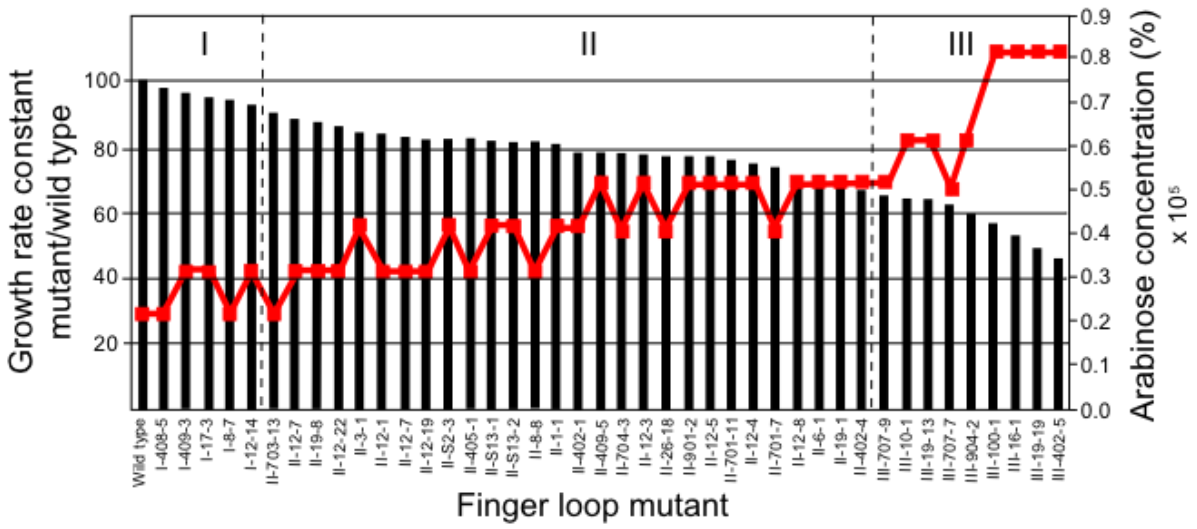

Fig. 3. Phenotypic classification of finger loop mutants. Growth rates of each mutant were calculated as a ratio of the mutant growth rate constant (GRC) /wild type growth rate constant, as indicated on the left ordinate. Mutants were partitioned into three functional groups (I, II and III) based upon these ratios. The concentration of arabinose required for growth of each mutant (right ordinate) is indicated by the red line. As shown, finger loop mutants with reduced growth rates in comparison with wild type, in general, required a higher concentration of arabinose to support growth.

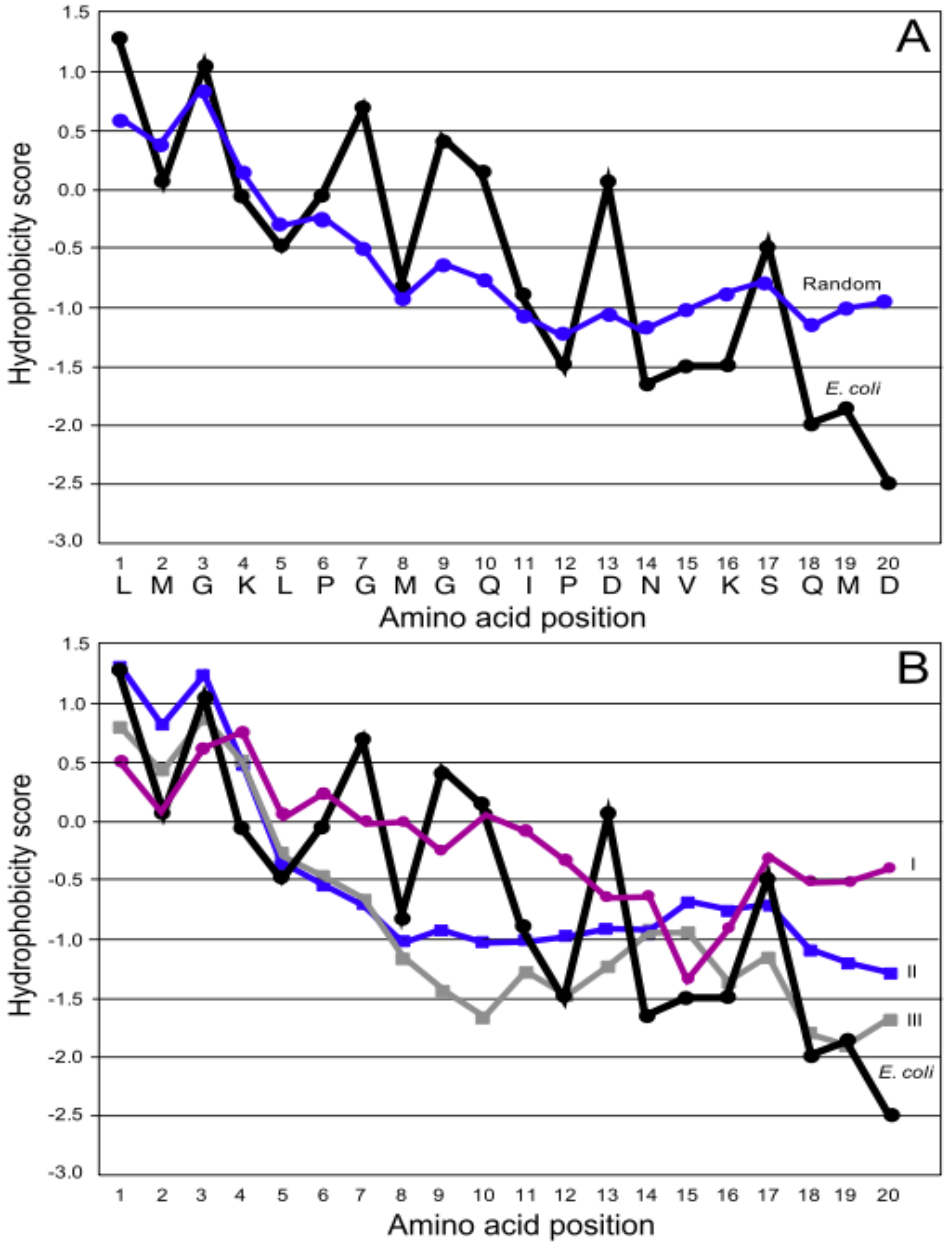

Fig. 4. Hydrophobicity plot of finger loop domain. (A) Comparison of the hydrophobicity values of *E. coli* Ffh (black line) with the average hydrophobicity of the 42 complementing sequences (blue line). Also shown is the average hydrophobicity at each position of a truly random sequence (grey line). (B) Comparison of the average hydrophobicity of each mutant group (I, purple line; II, blue line; III, grey line) and the hydrophobicity of wild type Ffh finger loop domain (black line). Hydrophobicity was determined by ProtScale using the Kyte-Doolittle scale.

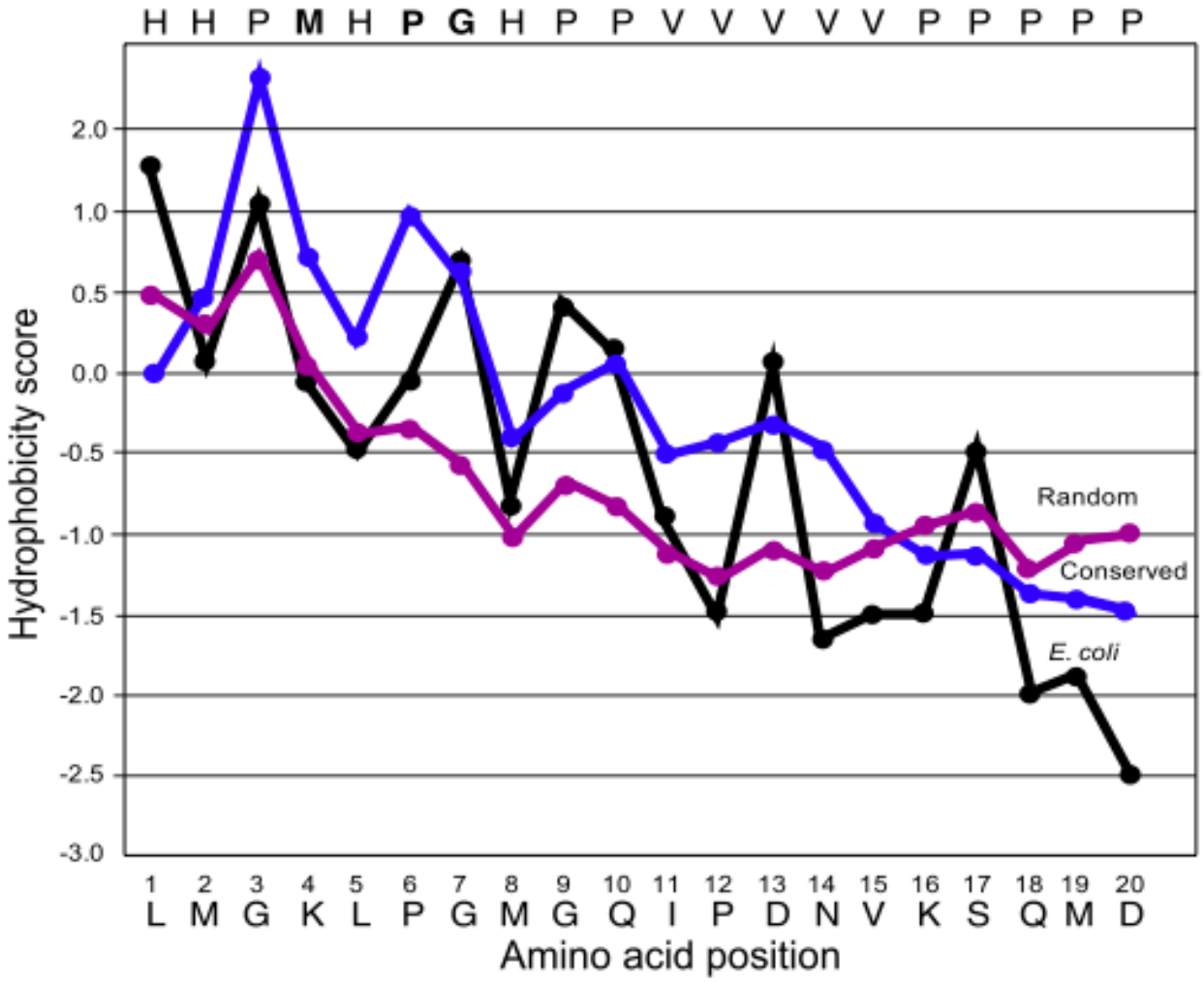

Fig. 5. Hydrophobicity plot of finger loop domain. Comparison of the hydrophobicity values of the finger loop domain from a subset of multiple species representing all three domains of life (blue line) with the average hydrophobicity of the 42 complementing sequences (purple line). The hydrophobicity of the finger loop domain from *E. coli* Ffh is also shown (black line). A consensus sequence displaying the conserved physicochemical properties or conserved residues for each amino acid position are indicated at the top of the graph, where H=Hydrophobic, P=Polar and V=Variable. Jalview 2.4.0 and WebLogo were used to determine the consensus sequence and hydrophobicity was determined by ProtScale using the Kyte-Doolittle scale.

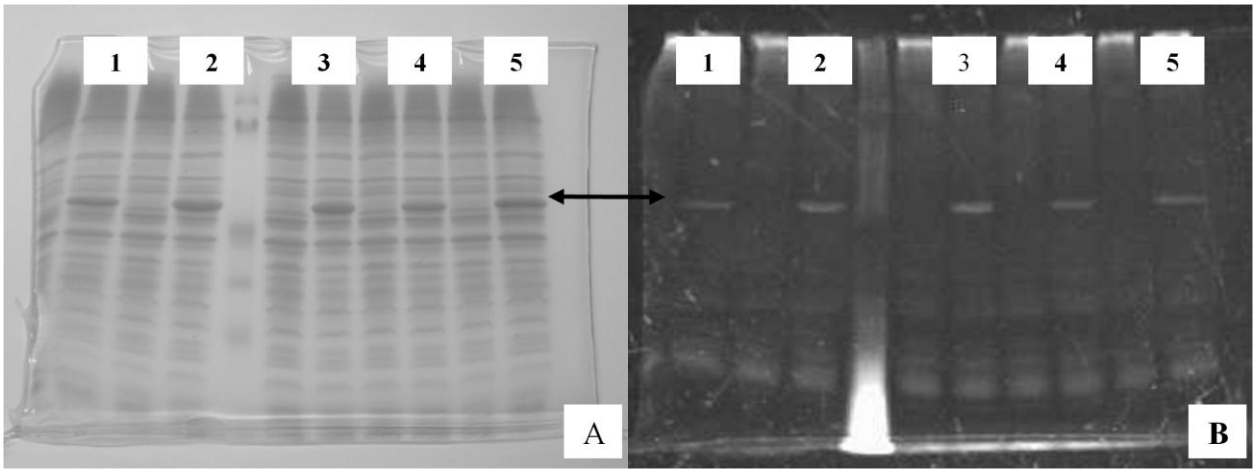

Fig. 6. Expression of finger loop mutants *in vivo*. (A) SDS-PAGE of total cell lysates expressing wild type Ffh (lane 1) and finger loop mutants L350A/M351A (lane 2), L350S/ M351S) (lane 3), D362L/N363Q/K365L (lane 4) and P355A/G356A (lane 5), where expression of *ffh* was induced by L-arabinose. (B) Detection of Ffh by His-tag stain, as described in Materials and Methods.

Table 1: Strains and plasmids used in this study

Strain of plasmid	Relevant genotype or description	Source or Reference
***E. coli* strain**		
NEB5α	*fhuA2*Δ(*argF-lacZ*)*U169 phoA glnV44 80* Δ(*lacZ*)*M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17* (general cloning host)	New England Biolabs
MC4100	F$^-$ *araD139*, Δ(*argF-lac*)U169, *rspL150*, *relA1*, *flbB5301*, *fruA25*, *deoC1*, *ptsF25* e14-	Lab collection
PMI105	MC4100, *ara*$^+$, *ffh*::*kan*1, Δ *recA-srl*)306 *srl-301*::Tn*10-84*, pFfhTS29 (Cam^R)	Lab collection
SLD106	PMI105, *ffh*::*kan*1, pFfhTS-Spc (Spc^R) (Tet^R)	This study
ECF529	Δ*araBAD*, Δ*rhaBAD*, Δ*araFGH*, Δ*araE*, *rrnB*PI(CTC-AGA)-*lacYA177C*	(3)
XLU102	ECF529, *bla*:: Δ*kan*	Lab collection
SLD108	XLU102, *ffh*::*kan*1, p*ffh*TS-Spc	This study
Plasmid		
p*ffh*TS-Spc	pSC101ts, *ffh*$^+$, *spc* (Spc^r)	Lab collection
pBAD*ffh*6x	Vector for expression of *ffh* under P_{araBAD} control, *araC*, *bla* (Amp^R)	(39)
pBAD*ffh*N6x	pBAD*ffh*6x (*Nhe*I)	This study
pBAD*ffh*N6xΔFL	pBAD*ffh*6x, P_{araBAD}-*ffh* ΔFL(*Nhe*I)	This study
pBAD*ffh*N6xFL*	pBAD*ffh*N6xΔFL plasmid series with random sequence insertions shown in Table 3.	This study
pBAD*ffh*N6xFL-PG-AA	pBAD*ffh*6x, PG to AA mutations	This study

Table 1: (continued)

Strain of plasmid	Relevant genotype or description	Source or Reference
Plasmid		
pBAD*ffh*N6xFL-LM-AA	pBAD*ffh*6x, LM to AA mutations	This study
pBAD*ffh*N6xFL-LM-SS	pBAD*ffh*6x, LM to SS mutations	This study
pBAD*ffh*N6xFL- DNK-LQL	pBAD*ffh*6x, D-L,N-Q,K-L mutations	This study

Table 2: PCR primers and oligonucleotides

Primer or oligonucleotide name	**Sequence (5'--3')**
FLrandom.S	GACTGACT**TCTAGA** $(NNK)_{20}$ **TCTAGA**TCAGTCAGTCAG
*ffh*FL2.S	TCTAGATCAGTCAGTCAG
*ffh*NFL.S	ATG**GCTAGC**AAAGTGCTGGTGCGTATGGAAGCC
*ffh*NFL.AS	CCCCCAGGCTTCCCTGGTCGC
*ffh*FL-LM-AA.S	A**GCTAGC**GCAGCCGGCAAGCTGCCGGGCATGGG
*ffh*FL-LM-SS.S	G**GCTAGC**TCCTCGGGCAAGCTGCCGGGCATGGG
*ffh*FLAS.AS	**GAGCTC**GCGACCAGGGAAG
*ffh*FL-DNK-LQL.S	CTA GTC TGA TGG GCA AGC TGC CGG GCA TGG GGC AGA TCC CGC TGC AGG TCC TGT CAC AGA TGC TGA
*ffh*FL-DNK-LQL.AS	CTA GTC AGC ATC TGT GAC AGG ACC TGC AGC GGG ATC TGC CCC ATG CCC GGC AGC TTG CCC ATC AGA
*ffh*FL-PG-AA.S	CTAGTCTGATGGGCAAGCTGGCCGCCATGGGCAGATCCCG GATAACGTCAAGTCACAGATGGACGATA
*ffh*FL-PG-AA.AS	CTAGTATCGTCCATCTGTGACTTGACGTTATCCGGGATCTG CCCCATGGCGGCCAGCTTGCCCATCAGA

Restriction enzyme sites are shown in bold

Table 3. Summary of finger loop mutant sequences and growth data

Finger loop mutant	Amino acid sequence	Growth rate constant[1]
Consensus Sequence[2]	HHP**M**HPGHPPVVVVVPPPPP	
Complementing Mutants		
GROUP I		
Wild Type	**LM**GKLPGMGQIPDNVKSQMD	1.28
I-408-5	HMWPGLLCRYASGNVTDVVI	1.25
I-409-3	KLAKTWDVAMNLEGSAGAVE	1.24
I-17-3	VVKLAQYKGVRVMESTEHNG	1.22
I-87	ILPLLPTRTILHQRSNPISD	1.22
I-12-14	QLHQILMPNTPLPPSTTHQQ	1.19
GROUP II		
II-703-13	LLAYMPSGHFMMRHVQGERE	1.15
II-21-7	LMKLRNTRIQSNQTITLHHL	1.13
II-19-8	IIQKLPHQMSTIQHIIPPPN	1.12
II-12-22	LCNWIVTHGLSRKGGAIQTE	1.10
II-3-1	FLAARSGNKSIPLSLRSEGG	1.10
II-12-1	VLAYLPYVSGMQSTGVWFGE	1.08
II-12-7	LFNFRESTSKKEAEGTTVPD	1.06
II-12-19	VIMEYGQMLAGTANVMSETQ	1.06
II-S2-3	IGELINQRMTGIYLHSCQRE	1.06
II-405-1	LLPKPHRHPLTPPTKHPISQ	1.06
II-S13-1	LMSWLRPFRARKGAHGFGSE	1.06

Table 3. (continued)

Finger loop mutant	Amino acid sequence	Growth rate constant[1]
GROUP II		
II-S13-2	LINRIQPPHKQTPPSQIHQN	1.05
II-8-8	VMELYQGLGGTRPDPRDSDQ	1.05
II-1-1	LLLPSMRLRTPKMRTTIPTP	1.04
II-402-1	LSIFLGSKLRFDQSDLFPDE	1.03
II-409-5	LISHTHTQHLLSTTPIRPQP	1.02
II-704-3	DLWNLMAAQGTKRRANRRDK	1.01
II-12-3	IIPLLPLPNTRPSIRNPQPT	1.01
II-26-18	ILELFAIEPKRMKRGTERCN	1.00
II-901-2	YAGHWAPARSASEKLLCVKD	1.00
II-12-5	FIQSLDRRMADHRYVSTCDE	1.00
II-701-11	ILQLPPSSIRLNTKQLPPTP	0.98
II-12-4	WGVKLVRTPGGRFLEPEVEA	0.97
II-701-7	FMDQLIEDNSCRRQTQHRIS	0.96
II-12-8	LMSLLRPQHINMNPLTQHIS	0.93
II-6-1	MIITSGGGPQTRGSTSGECS	0.92
II-19-1	LMTHQMTHTPTNLRNTNPIT	0.92
II-402-4	LMPMTRTPHPIKTLSTLNNQ	0.90
GROUP III		
III-707-9	FLGRCMPRSSTGMDGSPDDR	0.85
III-10-1	FAFIGRLSPVRPLRTVSTGQ	0.82
III-19-13	VRGFLSCGKTTQTNCDWASE	0.82

Table 3. (continued)

Finger loop mutant	Amino acid sequence	Growth rate constant[1]
GROUP III		
III-707-9	FLGRCMPRSSTGMDGSPDDR	0.85
III-10-1	FAFIGRLSPVRPLRTVSTGQ	0.82
III-19-13	VRGFLSCGKTTQTNCDWASE	0.82
III-707-7	YFGSTVNIPNNEMGAMEPQK	0.80
III-904-2	LLNQTTPSRIHRSPLQRRQQ	0.77
III-100-1	LLPRTTNRRHVGRHWVEPGY	0.73
III-16-1	ALQTIGSNVGPEDFTKVDNQ	0.68
III-19-19	VLDMLVSNTQEMTRGLERSD	0.63
III-402-5	RLGVSNTQSSDNASPREHIN	0.59
Non-Complementing Mutants		
PG-AA	LMGKLAAMGQIPDNVKSQM	
LM-AA	AAGKLPGMGQIPDNVKSQM	
LM-SS	SSGKLPGMGQIPDNVKSQM	
DNK-LQL	LMGKLPGMGQIPLQVLSQM	

[1]Growth rate constants for each mutant were determined as described in Materials and Methods. Mutants were classified as described in the legend to Fig 4. Hydrophobic amino acids Leu-Met in the wild type finger loop are shown in bold, while the conserved Pro-Gly is underlined in the wild type sequence.

[2]A consensus sequence showing the conserved physicochemical properties or conserved residues for each amino acid position in the finger loop are indicated at the top of the table where conserved Methionine (M) is shown in bold, conserved Pro-Gly is underlined, H=Hydrophobic, P=Polar and V=Variable. Jalview 2.4.0 and WebLogo were used to determine the consensus sequence and hydrophobicity was determined by ProtScale using the Kyte-Doolittle scale.

REFERENCES

1. **Andersen, E. S., M. A. Rosenblad, N. Larsen, J. C. Westergaard, J. Burks, I. K. Wower, J. Wower, J. Gorodkin, T. Samuelsson, and C. Zwieb.** 2006. The tmRDB and SRPDB resources. Nucleic Acids Res. **34:**D163-8.
2. **Batey, R. T., R. P. Rambo, L. Lucast, B. Rha, and J. A. Doudna.** 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science **287:**1232-9.
3. **Bowers, L. M., K. Lapoint, L. Anthony, A. Pluciennik, and M. Filutowicz.** 2004. Bacterial expression system with tightly regulated gene expression and plasmid copy number. Gene **340:**11-8.
4. **Brown, S., and M. J. Fournier.** 1984. The 4.5S RNA gene of *Escherichia coli* is essential for cell growth. J. Mol. Biol. **178:**533-50.
5. **Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton.** 2004. The Jalview Java alignment editor. Bioinformatics **20:**426-7.
6. **Clemons, W. M., Jr., K. Gowda, S. D. Black, C. Zwieb, and V. Ramakrishnan.** 1999. Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 Å resolution: evidence for the mechanism of signal peptide binding. J. Mol. Biol. **292:**697-705.
7. **Cleverley, R. M., N. Zheng, and L. M. Gierasch.** 2001. The cost of exposing a hydrophobic loop and implications for the functional role of 4.5 S RNA in the *Escherichia coli* signal recognition particle. J. Biol. Chem. **276:**19327-31.
8. **Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner.** 2004. WebLogo: a sequence logo generator. Genome Res. **14:**1188-90.
9. **Datsenko, K. A., and B. L. Wanner.** 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc. Natl. Acad. Sci. USA **97:**6640-5.
10. **de Gier, J. W., P. Mansournia, Q. A. Valent, G. J. Phillips, J. Luirink, and G. von Heijne.** 1996. Assembly of a cytoplasmic membrane protein in *Escherichia coli* is dependent on the signal recognition particle. FEBS Lett. **399:**307-9.
11. **Doudna, J. A., and R. T. Batey.** 2004. Structural insights into the signal recognition particle. Annu. Rev. Biochem. **73:**539-57.
12. **Driessen, A. J., and N. Nouwen.** 2008. Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. **77:**643-67.
13. **Dube, D. K., M. E. Black, K. M. Munir, and L. A. Loeb.** 1993. Selection of new biologically active molecules from random nucleotide sequences. Gene **137:**41-7.
14. **Egea, P. F., R. M. Stroud, and P. Walter.** 2005. Targeting proteins to membranes: structure of the signal recognition particle. Curr. Opin. Struct. Biol. **15:**213-20.
15. **Eisenberg, D., R. M. Weiss, and T. C. Terwilliger.** 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. USA **81:**140-4.
16. **Engelman, D. M., T. A. Steitz, and A. Goldman.** 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Biophys. Chem. **15:**321-53.
17. **Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch.** 2005. Protein identification and analysis tools on the ExPASy server, p. 571-607. *In* J. M. Walker (ed.), The Proteomics Protocols Handbook. Humana Press.

18. **Geyer, C. R., and R. Brent.** 2000. Selection of genetic agents from random peptide aptamer expression libraries. Methods Enzymol. **328:**171-208.
19. **Gill, D. R., and G. P. Salmond.** 1987. The *Escherichia coli* cell division proteins FtsY, FtsE and FtsX are inner membrane-associated. Mol. Gen. Genet. **210:**504-8.
20. **High, S., and B. Dobberstein.** 1991. The signal sequence interacts with the methionine-rich domain of the 54-kD protein of signal recognition particle. J. Cell. Biol. **113:**229-33.
21. **Hopp, T. P., and K. R. Woods.** 1983. A computer program for predicting protein antigenic determinants. Mol. Immunol. **20:**483-9.
22. **Ilangovan, U., S. H. Bhuiyan, C. S. Hinck, J. T. Hoyle, O. N. Pakhomova, C. Zwieb, and A. P. Hinck.** 2008. *A. fulgidus* SRP54 M-domain. J. Biomol. NMR **41:**241-8.
23. **Janda, C. Y., J. Li, C. Oubridge, H. Hernandez, C. V. Robinson, and K. Nagai.** 2010. Recognition of a signal peptide by the signal recognition particle. Nature.
24. **Keenan, R. J., D. M. Freymann, R. M. Stroud, and P. Walter.** 2001. The signal recognition particle. Annu. Rev. Biochem. **70:**755-75.
25. **Keenan, R. J., D. M. Freymann, P. Walter, and R. M. Stroud.** 1998. Crystal structure of the signal sequence binding subunit of the signal recognition particle. Cell **94:**181-91.
26. **Krieg, U. C., P. Walter, and A. E. Johnson.** 1986. Photocrosslinking of the signal sequence of nascent preprolactin to the 54-kilodalton polypeptide of the signal recognition particle. Proc. Natl. Acad. Sci. USA **83:**8604-8.
27. **Kurzchalia, T. V., M. Wiedmann, A. S. Girshovich, E. S. Bochkareva, H. Bielka, and T. A. Rapoport.** 1986. The signal sequence of nascent preprolactin interacts with the 54K polypeptide of the signal recognition particle. Nature **320:**634-6.
28. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157:**105-32.
29. **Luirink, J., G. von Heijne, E. Houben, and J. W. de Gier.** 2005. Biogenesis of inner membrane proteins in *Escherichia coli*. Annu. Rev. Microbiol. **59:**329-55.
30. **Lund, O., M. Nielsen, C. Lundegaard, and P. Worning.** 2002. Presented at the Critical Assessment of Protein Structure Prediction (CASP), Asilomar Conference Grounds, Pacific Grove, CA.
31. **Lutcke, H., S. High, K. Romisch, A. J. Ashford, and B. Dobberstein.** 1992. The methionine-rich domain of the 54 kDa subunit of signal recognition particle is sufficient for the interaction with signal sequences. Embo. J. **11:**1543-51.
32. **Macfarlane, J., and M. Muller.** 1995. The functional integration of a polytopic membrane protein of *Escherichia coli* is dependent on the bacterial signal-recognition particle. Eur. J. Biochem. **233:**766-71.
33. **Montoya, G., C. Svensson, J. Luirink, and I. Sinning.** 1997. Crystal structure of the NG domain from the signal-recognition particle receptor FtsY. Nature **385:**365-8.
34. **Morgan-Kiss, R. M., C. Wadler, and J. E. Cronan, Jr.** 2002. Long-term and homogeneous regulation of the *Escherichia coli* araBAD promoter by use of a lactose transporter of relaxed specificity. Proc. Natl. Acad. Sci. USA **99:**7373-7.
35. **Nagai, K., C. Oubridge, A. Kuglstatter, E. Menichelli, C. Isel, and L. Jovine.** 2003. Structure, function and evolution of the signal recognition particle. Embo. J. **22:**3479-85.
36. **Oh, D. B., G. S. Yi, S. W. Chi, and H. Kim.** 1996. Structure of a methionine-rich segment of *Escherichia coli* Ffh protein. FEBS Lett. **395:**160-4.
37. **Palzkill, T., Q. Q. Le, A. Wong, and D. Botstein.** 1994. Selection of functional signal peptide cleavage sites from a library of random sequences. J. Bacteriol. **176:**563-8.

38. **Patel, P. H., and L. A. Loeb.** 2000. DNA polymerase active site is highly mutable: evolutionary consequences. Proc. Natl. Acad. Sci. USA **97:**5095-100.
39. **Peterson, J. M., and G. J. Phillips.** 2008. Characterization of conserved bases in 4.5S RNA of *Escherichia coli* by construction of new F' factors. J. Bacteriol. **190:**7709-18.
40. **Phillips, G. J.** 1999. New cloning vectors with temperature-sensitive replication. Plasmid **41:**78-81.
41. **Phillips, G. J., and T. J. Silhavy.** 1992. The *E. coli ffh* gene is necessary for viability and efficient protein export. Nature **359:**744-6.
42. **Poritz, M. A., H. D. Bernstein, K. Strub, D. Zopf, H. Wilhelm, and P. Walter.** 1990. An *E. coli* ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. Science **250:**1111-7.
43. **Ribes, V., K. Romisch, A. Giner, B. Dobberstein, and D. Tollervey.** 1990. *E. coli* 4.5S RNA is part of a ribonucleoprotein particle that has properties related to signal recognition particle. Cell **63:**591-600.
44. **Romisch, K., J. Webb, K. Lingelbach, H. Gausepohl, and B. Dobberstein.** 1990. The 54-kD protein of signal recognition particle contains a methionine-rich RNA binding domain. J. Cell. Biol. **111:**1793-802.
45. **Rosendal, K. R., K. Wild, G. Montoya, and I. Sinning.** 2003. Crystal structure of the complete core of archaeal signal recognition particle and implications for interdomain communication. Proc. Natl. Acad. Sci. USA **100:**14701-6.
46. **Rost, B., G. Yachdav, and J. Liu.** 2004. The PredictProtein server. Nucleic Acids Res. **32:**W321-6.
47. **Schneider, T. D., and R. M. Stephens.** 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18:**6097-100.
48. **Skandalis, A., and L. A. Loeb.** 2001. Enzymatic properties of rat DNA polymerase beta mutants obtained by randomized mutagenesis. Nucleic Acids Res. **29:**2418-26.
49. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673-80.
50. **Valent, Q. A., D. A. Kendall, S. High, R. Kusters, B. Oudega, and J. Luirink.** 1995. Early events in preprotein recognition in *E. coli*: interaction of SRP and trigger factor with nascent polypeptides. Embo. J. **14:**5494-505.
51. **von Heijne, G.** 1985. Signal sequences. The limits of variation. J. Mol. Biol. **184:**99-105.
52. **Walter, P., and G. Blobel.** 1980. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. Proc. Natl. Acad. Sci. USA **77:**7112-6.
53. **Walter, P., and G. Blobel.** 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. Nature **299:**691-8.
54. **Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones.** 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J .Mol. Biol. **337:**635-45.
55. **Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton.** 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **25:**1189-91.

56. **Zheng, N., and L. M. Gierasch.** 1997. Domain interactions in *E. coli* SRP: stabilization of M domain by RNA is required for effective signal sequence modulation of NG domain. Mol. Cell **1:**79-87.
57. **Zopf, D., H. D. Bernstein, A. E. Johnson, and P. Walter.** 1990. The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be crosslinked to a signal sequence. Embo. J. **9:**4511-7.

CHAPTER 3. Methionine Bristles in the Signal Sequence Binding Domain of Ffh are not Required for Function of the *Escherichia coli* Signal Recognition Particle

A paper to be submitted to *PLoS One*

Stacy S. Duncan and Gregory J. Phillips

ABSTRACT

The signal recognition particle (SRP) is a ribonucleoprotein complex important for targeting proteins to the cytoplasmic membrane in all living cells. Ffh, the sole protein component of the SRP in *E. coli*, binds to signal peptides via interaction with a methionine rich carboxy-terminus known as the M-domain. It was originally proposed that signal peptides bind the M-domain via interaction with "methionine bristles" formed by side chains of this amino acid located primarily on the same face of four predicted alpha helices. Despite the appeal of this model, no direct test of the importance of methionine in SRP function has been made. We used both phylogenetic sequence comparisons and mutagenesis, including replacing methionine residues with several other amino acids to test the methionine bristle hypothesis. In addition to methionine, we found leucine and isolecucine are also highly abundant in the M-domain, actually surpassing methionine in prevalence in hyperthermophilic microorganisms. By directly testing SRP function in a variety of replacement mutants we unexpectedly observed that the requirement for methionine does not correlate with amino acid conservation. Substitution of all the methionines in the alpha M4 helix and the extreme carboxy-terminus of Ffh with leucine

failed to functionally replace methionine, while isoleucine only poorly replaced methionine. As predicted, similar results were observed when charged or non-hydrophobic amino acids were used to replace methionine. However, other amino acids, although far less conserved in Ffh/SRP54, functioned well as methionine replacements. While phenylalanine, tyrosine and tryptophan all supported SRP function to varying degrees, valine restored SRP activity to nearly wild type levels. Furthermore, valine could replace all but one highly conserved methionine within the entire M-domain. This single, essential methionine residue is likely important for the dense packing of alpha helices within the M-domain, and does not appear to directly participate in signal peptide binding. These results reveal a surprising degree of adaptability in substrate binding and helps explain how the SRP can target many different signal peptides and transmembrane domains. Further characterization of these mutants will reveal new insights into how the structure of the M-domain contributes to SRP function.

INTRODUCTION

The signal recognition particle (SRP) is a universally conserved ribonucleoprotein complex important for targeting polypeptides to cellular membranes in eukaryotes, archaea and bacteria (10-12, 17). The eukaryotic SRP is comprised of a 7S RNA and six proteins (39, 40), while in bacteria and archaea it is a simpler complex. In *E. coli*, for example, the SRP includes the Ffh (54-homolog) protein in complex with 4.5S RNA (3, 30, 32, 33). In all organisms, SRP mediated protein localization involves similar mechanisms. As nascent polypeptides emerge from the ribosome, the SRP binds to hydrophobic signal sequences or transmembrane domains on nascent polypeptides via interaction with the Ffh (SRP54) protein. Subsequently, the SRP-ribosome-peptide chain complex associates with the SRP receptor (FtsY in *E. coli*) where the target protein is inserted into or across the membrane in a reaction that requires GTP hydrolysis by both components of the SRP and FtsY herterodimer (10-12, 17). Consistent with its role in protein targeting, the structural genes encoding components of the SRP pathway, i.e., *ffh*, *ffs* and *fts*Y, are all essential for cell viability in *E. coli* (6, 19, 29).

Ffh is comprised of three domains, the amino-terminal N-domain, GTP binding domain (G-domain) and the carboxy-terminal, methionine rich M-domain (2, 7, 13-15, 18, 25, 35). Biochemical and structural analysis (2, 7, 13-15, 18, 25, 35) have provided insights into how these domains function in membrane protein localization. The amino terminal region of Ffh, comprised of the N and G domains, mediate the binding of SRP with its receptor. Ffh and FtsY share sequence and structural homology in their N and G domains (24), which is consistent with their interaction to form a functional GTPase (21, 22, 27, 31, 36, 42). The M-domain contains binding sites for 4.5S RNA and hydrophobic signal sequences (20, 34, 43).

Insights as to how the SRP appears to bind its membrane protein cargo initially came from sequence analysis of the *ffh* gene cloned from *E. coli* (3). Upon observing that the carboxy terminal domain is rich in methionines, Bernstein *et al.* predicted that the unique characteristics of this amino acid are important for binding signal sequences. Since the side chains of methionines are hydrophobic, yet retain flexibility, they likely form the signal sequence binding region of the protein. In this model, properly positioned R-groups protrude outward to form "bristles" that allow signal sequences to bind through hydrophobic interactions. Because the SRP must be able to bind a variety of signal sequences, it was thought that the bristles would allow for adaptability of the region (3).

Indeed, crystal structures of the Ffh protein from multiple sources generally support the "methionine bristle" model (2, 7, 13-15, 18, 25, 35). In all cases, the M-domain is composed of multiple α-helices that, along with a flexible loop known as the finger loop, form a hydrophobic binding pocket for membrane proteins that is exposed on the surface of the protein. The binding region can accommodate hydrophobic transmembrane domains and signal sequences of exported proteins through interaction with the side chains of methionine residues, as well as other hydrophobic amino acids found in the carboxy-terminus of Ffh. Figure 1A shows the structure of the M-domain of *Thermus aquaticus* (18). As in most Ffh proteins, the M-domain consists of four α helices (αM1-4), arranged such that αM1, αM2 and αM4 along with the finger loop, which connects αM1 and αM2, form the hydrophobic binding groove. Structural determination also shows that this groove is lined with the side chains of several hydrophobic residues, including a preponderance of methionine. More recently, the crystal structure of an Ffh-signal peptide complex was determined (15). As predicted from previously determined structures (2, 7, 13-15, 18, 25, 35), the crystal structure revealed the model signal peptide indeed binds in the

hydrophobic groove formed by αM1, αM2 and αM4 (Fig. 1B). Furthermore, while the signal peptide makes significant contacts with αM4, the finger loop appears to form a "lid" over the hydrophobic binding groove (15). A similar structure was predicted to be formed by the *T. aquaticus* M-domain (18), with the exception that many of the methionine residues were substituted with other hydrophobic amino acids including leucine, isoleucine and phenylalanine. The increase in thermal motion of the amino acid side chains at the higher growth temperatures of *T. aquaticus* appears to compensate for the reduced flexibility of the side chains of leucine, isoleucine and phenylalanine (18).

Despite the predicted importance of methionine residues for Ffh function, there have been no direct tests of the "methionine bristle" hypothesis. To address this, we have performed a systematic analysis of the Ffh M-domain both by comparative analysis and by mutagenesis. Mutants were generated by replacing the ten methionine residues in αM4 and the extreme carboxy-terminal region of Ffh, representing half of the total number residues of this amino acid in the M-domain, with several other amino acids. By characterizing each mutant for viability, growth, and SRP function we were able to identify specific amino acid substitutions that can functionally replace methionine, hence providing new insights into signal peptide binding by Ffh.

MATERIALS AND METHODS

Bacterial strains and plasmids. The strains and plasmids used in this study are shown in Table 1.

Reagents. Restriction enzymes and other enzymes used for recombinant DNA were obtained from New England Biolabs (Ispsich, MA) and Fermentas Life Sciences (Glen Burnie, MD).

PCR primers and synthetic genes were obtained from Integrated DNA Technologies (Coralville, IA). Growth medium were obtained from Difco (Detroit, MI). All antibiotics and other chemicals were obtained from Sigma Chemical Co. (St. Louis, MO).

Construction of Ffh M-domain mutant alleles. Each M-domain mutant (Table 2) was constructed by gene synthesis (Integrated DNA Technologies, Coralville IA). DNA constructs corresponding to the region of *ffh* encoding the αM4 helix (Fig. 1) and the unstructured carboxy terminus of the protein were obtained from IDT (Coralville Iowa). Each 126-bp construct, was synthesized so that all of the base triplets specifying methionine were replaced by alanine, cysteine, glutamate, isoleucine, leucine, phenylalanine, tryptophan, tyrosine or valine. Each construct was digested with *Xmn*I-*Sac*I and used to replace a similar segment from pSLD-*ffh*10 (Table 1). Plasmid DNA was isolated using a Qiagen miniprep kit (Qiagen, Valencia, CA) and all constructs were confirmed by DNA sequencing (DNA Sequencing and Synthesis Facility at Iowa State University). Two additional synthetic constructs were also designed by replacing every triplet encoding methionine in the entire M-domain with either valine or a combination of valine and methionine (Table 1). These segments of *ffh* were introduced to pSLD-*ffh*10 using *Pvu*II-*Sac*I.

To construct plasmids expressing *ffh* under *lac* control, an *Age*I and *Aat*II segment was replaced with the corresponding region from pSLD-*ffh*10 and 15.

Western blot analysis. To facilitate detection of Ffh, DNA encoding a cMyc epitope tag was positioned onto the end of each *ffh* construct. A construct carrying a DNA fragment encoding the cMyc epitope tag (EQKLISEEDL) was obtained from IDT (Coralville Iowa). The cMyc epitope tag was introduced to the pSLD-*ffh* plasmids at the 5' of *ffh* using *FspA*I and *Sal*I.

Complementation tests. The pSLD-*ffh* plasmids were transformed into SLD106 (Table 1) as described previously (29). Transformants were plated in duplicate on LB agar plates supplemented with ampicillin (Amp) (100µg/ml) and incubated at 30°C and 42°C. Amp^R transformants that grew at 42°C were restreaked on LB agar plus spectinomycin (Spc) (100µg/ml) plates and incubated at 30°C to test for Spc^s, indicating loss of the temperature-sensitive plasmid pTS-Spc-*ffh*$^+$. To test the ability of pLac-*ffh* derivative plasmids to complement transformants were plated in duplicate on LB agar plates supplemented with Amp (100µg/ml) plus isopropyl-β-D-thiogalactopyranoside (IPTG) (0.1 mM) and incubated at 30°C and 42°C. Amp^R transformants were restreaked on LB plus Spc plates and incubated at 30°C to test for Spc^s.

Construction of F' plasmids. Recombineering was used to construct new F' plasmids, similar to that described by Peterson and Phillips (29). For this, a gene cassette encoding chloramphenicol resistance (Cam^R) was first introduced to a unique *Xba*I restriction site at the 3' end of *ffh* on the plasmids pSLD-*ffh*10, pSLD-*ffh*14, pSLD-*ffh*16, pSLD-*ffh*17, pSLD-*ffh*18 and pSLD-*ffh*19. PCR was used to amplify *ffh-cam* using primers *ffh-lacA*.S2 and *ffh-lacA*.AS (5'-ATTGCACCCAACGTTACTCTTTCCGTTACGGGACACCCTGTACACGCAAGATTCCGAATACCGCAAG-3' and 5'-CGCCACGACGTTTGGTGGAATGTCTTTTGTGACGATACTACCCGCCAGCTATGACCATGATTACGCC-3') using cycling conditions of: 2 min at 94°, 30 cycles of (30 sec at 94°, 2 min at 55°, 1 min at 72°), 10 min at 72°. Each PCR product was gel purified using the Qiagen QIAquick Gel Extraction kit (Qiagen, Valencia, CA) and was introduced to F'*lac proA*$^+$*B*$^+$ by homologous recombination at *lacA* using recombineering. Recombineering was carried out by transferring 0.05 ml of overnight cultures of CSH100 F'*lac proA*$^+$*B*$^+$ to 5 ml of fresh LB and grown at 30°C with continuous shaking until an OD_{600} of 0.6

was reached. To induce the Red recombinase genes the culture was shifted to 42°C with continued shaking for 15 minutes. Cultures were immediately placed in an ice water slurry and swirled for 5 minutes. Cells were pelleted at 4°C and washed four times with ice cold, sterile, nanopure water and were subsequently resuspended in 50 μl of water. ~1 μg of each PCR product was electroporated into CSH100 F'*lac proA*$^{+}$*B*$^{+}$. Cam^{R} recombinants were selected at 30°C on LB plus Cam (12.5 μg/ml) plates. Conjugation was then used to move each of the new F' factors into the recipient strain SLD106 (Table 1). The donors and recipient were patched together on an LB agar plate and incubated at 30^{o}C for 6 hours. Transconjugates were selected by streaking cells onto LB agar plates supplemented with Cam (12.5 μg/ml) and Kan (30 μg/ml) to counterselect against the donor strains.

Growth rate measurements. To compare growth rates of the *ffh* mutants expressed at different levels and at different temperatures, Spc^{s} transformants were restreaked on LB agar plates supplemented with Cam (20μg/ml), when testing pSLD-*ffh* mutants, or Amp (100μg/ml) plus IPTG (0.1 mM), when testing pLac-*ffh* mutants, and incubated at 26^{o}C, 30^{o}C, 37^{o}C, and 42^{o}C.

To compare growth of the mutants in single copy, a single colony from each of the SLD106 F' transformants were inoculated into 50 ml LB plus Cam (20μg/ml) and grown with aeration at 30^{o}C, 37^{o}C or 42^{o}C. Samples were removed at half hour intervals from the cultures for measurement of OD_{600}.

SRP dependent localization of FtsQ assay. SRP dependent localization of a modified inner membrane protein FtsQ was detected using the biotinylation assay described previously (26, 38). Each SLD106 derivative mutant carrying wild-type *ffh* or the *ffh*α4M→I, *ffh*α4M→W, *ffh*α4M→F, *ffh*α4M→Y and *ffh*α4M→V alleles was transformed with plasmid pBAD*ftsQ*-V5-PSBT (29). Cells were grown at 37^{o}C in 5 ml of LB broth supplemented with Cam (12.5 μg/ml)

to mid-logarithmic phase and subsequently induced with 0.02% L-arabinose. Upon reaching an OD_{600} of ~0.6, a 2 ml volume of the cells were pelleted by centrifugation at 13,000 rpm in a microcentrifuge. The pellets were resuspended in 100µl of nanopure water. Subsequently, 40µl of the samples were mixed with 38µl of Laemmli sample buffer (Bio Rad, Hercules, CA) and 2µl of β-mercaptoethanol, boiled for 5 min, and then recentrifuged and prepared for SDS-PAGE analysis using Pierce 8-16% Precise Protein Gels (Pierce, Rockford, IL). Proteins were transferred to a nitrocellulose membrane and probed for FtsQ using rabbit anti-V5 antibody (Bethyl Inc., Montgomery, TX) and goat anti-rabbit-horseradish peroxidase conjugate (Pierce, Rockford, IL). As described by Park *et al*. (26), biotinylated FtsQ was detected using streptavidin-horseradish peroxidase conjugate (Pierce, Rockford, IL). The SuperSignal West Femto substrate kit (Pierce, Rockford, IL) was used for protein visualization and imaged on a ChemiImager 5500 (Alpha Innotech, San Leandro, CA).

β-galactosidase assays. The plasmid pBAD-*ftsQ-lacZ* (Table 1) was constructed by inserting *lacZ* (generated by PCR) into the unique *Kpn*I site in *ftsQ* on pBAD*ftsQ*-V5-PSBT as to fuse *ftsQ* in frame with *lacZ*, pBAD-*ftsQ-lacZ* was transformed into strains SLD106 containing the pSLD-*ffh* plasmids (Table 1). Transformants were plated on LB agar plates supplemented with Cam (20µg/ml) plus Amp (100µg/ml) plus 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) (40 ug/ml) and incubated at 37^{o}C and 42^{o}C (in the case of mutant *ffh*α4M→I, which only grows at 42^{o}C). β-galactosidase assays were performed on overnight cultures grown in the absence of arabinose, as described by (23).

Sequence comparisons. SRP54/ Ffh sequences were downloaded from the NCBI database (1). Multiple sequence alignments and consensus sequences for 20 hyperthermophiles (16 archaea and 4 bacteria), 46 thermophiles (8 archaea and 38 bacteria), 11 psychrophiles (2 archaea and 9

bacteria) and 63 mesophiles (2 archaea and 61 bacteria) were generated using Jalview (41) and Weblogo (8).

RESULTS

Distribution of methionine in the M-domain of Ffh. Upon determining the crystal structure of the Ffh M-domain from *Thermus aquaticus*, Keenan *et al*. (18) predicted that several conserved methionine residues are substituted in this thermophile with other hydrophobic amino acids with less flexible side chains. To test this hypothesis, using much larger data sets than that available to Keenan *et al*., we selected *ffh* sequences from microorganims representing a range of optimal growth temperatures (OGT), including hyperthermophiles (OGT $\geq$75^{o}C), thermophiles (40^{o}C-75^{o}C), mesophiles (25^{o}C-40^{o}C) and psychrophiles (10^{o}C-15^{o}C). As shown in Fig 3, while residues with less flexible side chains (e.g. leucine, phenylalanine, isoleucine and valine) were indeed substituted for methionines in the M-domain of hyperthermophiles and thermophiles, mesophiles and psycrophiles were comprised extensively of methionines.

While the positions of most of the methionines were highly variable throughout the M-domain, we observed three positions, (corresponding to residues 344, 383 and 426 of *E. coli* Ffh), were very highly conserved among all four groups of microorganisms. Expanding the comparison to include all three domains of life revealed that these three conserved positions were nearly invariant (data not shown), with position 383, located in αM3, the SRP RNA binding helix, being 100% conserved. Methionine residues corresponding to positions 344 and 426 of *E. coli* Ffh were 72% and 44% conserved in hyperthermophiles, 54% and 59% in thermophiles, 73% and 75% in mesophiles, and 100% and 91% in psychrophiles, respectively. These positions are highlighted in Fig. 1B and Fig.1D.

Construction of new Ffh M-domain mutants. Of the twenty methionines in the M-domain of *E. coli* Ffh, half are contained within the αM4 helix (positions 423, 426, 427, 430) and unstructured "tail" at the extreme carboxy-terminus (positions 435, 438, 439, 442, 445 and 446). As shown in Fig. 3, although methionines are clustered in this region of the M-domain, there is significant variability in amino acid composition of the unstructured tail. While methionines in the αM4 helix were almost exclusively substituted with hydrophobic residues, methionines found in the extreme carboxy-terminal tail were frequently substituted with polar and charged residues. To assess the importance of the 10 methionine residues for SRP function in *E. coli*, we replaced all 10 amino acids with residues containing different chemical properties. The replacements included hydrophobic amino acids with non-bulky side chains (alanine, isoleucine, leucine, and valine); hydrophobic, with bulky side chains (phenylalanine, tryptophan, and tyrosine); a charged side chain (glutamate); and the other hydrophobic, sulphur containing side chain (cysteine). As summarized in Table 1, we designated each of these mutant alleles as *ffh*α4M.

Initially, we expressed all mutants from a multiple copy number plasmid (Materials and Methods). The ability of each mutant to function as a component of the SRP was tested by transforming each plasmid construct into SLD106 (Table 1). This strain carries *ffh*::*kan*1 allele (26), which is complemented by wild type *ffh* carried on a plasmid that confers Spc^R and is temperature sensitive for replication. As a consequence, SLD106 grows at 42^{o}C only if it is provided with a functional copy of *ffh*. As shown in Fig. 4, we observed that mutants *ffh*α4M→A, *ffh*α4M→C *ffh*α4M→E and *ffh*α4M→L all failed to complement *ffh*::*kan1*. The remaining transformants were confirmed to be Spc^S, indicating loss of the temperature sensitive plasmid.

To quantify the expression of the *ffh*α4M mutants, we modified each of the pSLD-*ffh* plasmids (Table 1) by introducing a DNA fragment encoding the cMyc epitope tag (EQKLISEEDL) to the 5' of *ffh*. Complementation tests were repeated with each cMyc derivative with similar results (data not shown). Western blot analysis using antibody against the cMyc epitope revealed that all of the mutants expressed Ffh at levels comparable to the wild type protein (Fig. 5A).

Although the *ffh*α4M alleles were expressed from multiple copy (ColE1-deriviative) number plasmids (Table 1), we reasoned that further elevating the expression of the mutants that failed to complement could restore their function. To test this, we cloned *ffh*α4M→A, *ffh*α4M→C, *ffh*α4M→E and *ffh*α4M→L into a plasmid where their expression was under control of the *lac* promoter. We observed that while *ffh*α4M→A, *ffh*α4M→C and *ffh*α4M→E again yielded no colonies when transformed into SLD106 in the presence of IPTG, *ffh*α4M→L gave rise to small colonies. However, the colonies failed to form single colonies upon restreaking (data not shown).

Phenotypes of M-domain mutants expressed at physiological levels. To more accurately determine how well each allele functioned as a component of the SRP, we tested the ability of the five functional mutants to complement when expressed in single copy. Using the strategy described by Peterson and Phillips (28), we constructed a series of F' plasmids and repeated the complementation tests after introducing wild type *ffh* and each of the functional *ffh*α4M alleles (*ffh*α4M→F, *ffh*α4M→I, *ffh*α4M→V, *ffh*α4M→W, and *ffh*α4M→Y) into SLD106 in single copy. Following conjugation with SLD106, transconjugants were selected at 42^{o}C to test for complementation. Although the wild type control and the five functional *ffh*α4M alleles were able to complement *ffh::kan*1 in SLD106 in single copy, the mutants complemented with

different degrees of efficiency. Because several of the substitutions were made with amino acids that were more hydrophobic, with less flexible side chains than methionine, we determined if growth temperature could influence the results of complementation test results. Upon restreaking each of the SLD106 transformants at 30°C, 37°C, and 42°C we noted differences in growth, suggesting that the products of the alleles were not functionally equivalent. For example, as shown in Fig. 6A (Top row), *ffh*α4M→I did not grow at 30°C. To quantify differences in the growth rate of each mutant, strains were cultured at 30°C, 37°C and 42°C and growth rate constants were measured. As shown in Table 2, both *ffh*α4M→I and *ffh*α4M→W grew optimally at 42°C, whereas the remaining mutants displayed optimal growth at 37°C.

Ffh M-domain mutants affect SRP function. To directly test the effect of the amino acid substitutions on SRP function we monitored the localization of FtsQ, an inner membrane protein whose localization is SRP dependent (37, 38). For this, we expressed a derivative of FtsQ that includes the biotinylation domain of transcarboxylase from *P. shermanii* (PSBT). In wild type cells, membrane protein targeting is so efficient that the protein fails to be biotinylated. However, when SRP function is compromised, FtsQ is retained in the cytoplasm and becomes modified by biotinylation (26, 37, 38). For this assay we used pBAD*ftsQ*-V5-PSBT (16, 28). As shown in Fig. 7A, almost no biotinylated FtsQ was detected in the wild type control strain; however, various amounts were detected in the mutants, the greatest being observed in mutants *ffh*α4M→I, *ffh*α4M→W and *ffh*α4M→Y. The degree of biotinylation correlated directly with the growth rates of each mutant, with the greatest defects observed in mutants *ffh*α4M→I, *ffh*α4M→W and *ffh*α4M→Y (Fig. 7A lanes 3, 5 and 6).

β-galactosidase activity of Ffh M-domain mutants. As an independent means to test the effect of the functional *ffh*α4M mutants on SRP function, we measured β-galactosidase activity from

strains expressing an *ftsQ-lacZ* fusion encoding a hybrid protein where the signal sequence of *ftsQ* is fused to the amino terminus of *lacZ*. It is well established that the product of *lacZ,* encoding a cytoplasmic enzyme, exhibits low activity when localized to the membrane (4, 37, 38). Consequently, β-galactosidase activity is an indirect measurement of SRP function since activity is detected only when the protein is poorly targeted to the membrane. As shown in Fig. 7B, the amount of β-galactosidase activity determined for each *ffh*α4M mutant correlated with the results of the FtsQ biotinylation assays described above (Fig. 7A).

A truncated *ffh* mutant. Given the results that substitutions of up to 10 methionine residues in the C-terminus of Ffh still supports function of the SRP in *E. coli*, we further distinguished the importance between residues located in the αM4 helix and those in the unstructured extreme end of the protein. For this we introduced an amber nonsense mutation corresponding to position 436 of Ffh and tested the ability of this mutant to support growth of SLD106. The truncated *ffh* mutant was able to complement *ffh*::*kan1* in SLD106 when expressed in multiple copy; however, it failed to support cell viability when expressed in single copy. Western blotting confirmed the product of this allele was expressed at levels comparable to that of wild type (Fig. 5B).

Construction of a "valine thistle" mutant. Since substitution of valine for methionine in αM4 resulted in nearly wild type SRP activity (Fig. 6-Top row and Fig. 7), we next determined if replacement of every methionine within the Ffh M-domain was likewise functional. For this, we constructed the *ffh*M→Vα1-4 allele by substituting all 20 ATG triplets within the M-domain with GTG. Although the product of this mutant allele was expressed at levels comparable to wild type, as confirmed by Western blotting (Fig. 5B), it failed to complement in SLD106 (Fig. 6- Bottom row). As described above, we identified 3 highly conserved methionines at positions 344, 383 and 426 (Fig. 1D). To determine the importance of these residues, we modified the

*ffh*M→Vα1-4 mutant to include these 3 highly conserved methionines (*ffh*V→M(x3)). In contrast to the *ffh*M→Vα1-4 mutant, the product of *ffh*V→M(x3) supported growth of SLD106 at nearly wild type levels when expressed in both multiple copy number and in single copy (Fig. 6-Bottom row). Further characterization of this mutant and its effect on SRP function showed that localization of the *ftsQ-lacZ* gene product is slightly reduced, as indicated by elevated β-galactosidase activity (Fig. 7B).

DISCUSSION

The importance of methionine residues for signal sequence binding by Ffh was first proposed by inspection of the predicted primary sequence of the *E. coli* Ffh protein (3). Bernstein *et al*. (3) proposed that the flexible, hydrophobic side chains of methionine represented "bristles" that formed the basis for signal peptide recognition by the SRP. Since then, several Ffh/SRP54 proteins have been crystallized and confirm that the side chains of methionine and other hydrophobic amino acids line the signal peptide binding groove of the M-domain (Fig. 1) (2, 7, 13-15, 18, 25, 35). Keenan *et al*. (18) further proposed that leucine and isolecuine, could substitute for methionine in organisms whose higher optimal growth temperatures could compensate for the less flexible R groups of these amino acids.

Since these original proposals, much larger data sets have become available for analysis. We used multiple sequence alignment comparisons, to measure the prevalence of methionine in the M-domain of Ffh/SRP from a wide range of organisms, including microorganims with different optimal growth temperatures. As shown in Fig. 2, we found that the prevalence of methionine in the M-domain largely correlates with the optimal growth temperature of its host organism. While methionine comprises 11% of the amino acids in the M-domain from

mesophiles, including *E. coli*, the percentage increases to 13% in psycrophiles whose optimal growth temperature is <20^{o}C. As originally proposed by Keenan *et al*. (18) leucine and isoleucine replaced methionine in thermophiles (OGT range is 40^{o}C-75^{o}C), and even to a great extent in hyperthermophiles (OGT ≥75^{o}C).

While sequence comparisons suggested the importance of methionine to SRP function, particularly in mesophiles and psychrophiles, we directly tested this by constructing several M-domain variants and determining their ability to function in *E. coli*. We observed that of the 20 methionine residues in the *E. coli* Ffh M-domain, 50% were clustered in αM4 and the carboxy-terminal end of the protein (Fig. 1). Phylogenetic sequence analysis of the four groups of microorganisms revealed that methionines occur more frequently in αM4 with increased residue variability in the carboxy-terminus (Fig. 3). Recent structural analysis of SRP and a signal peptide (15) indicates that αM4 contributes a significant "platform" to the signal peptide binding groove (Fig. 1). Consequently, we simultaneously replaced all of the methionines in αM4 (4 residues) and the unstructured carboxy-terminal tail (6 residues) with amino acids differing in hydrophobicity, side chain flexibility and charge (Table 2). Surprisingly, the requirement for methionine is not as stringent as suggested from phylogenetic comparisons. The *ffh*α4M→L mutant was not viable, even when overexpressed (Fig. 4-Bottom row), and *ffh*α4M→I functioned well only when expression was elevated (Fig. 4-Top row). In contrast, the *ffh*α4M→V mutant supported growth near wild type levels when expressed in single and multiple copy (Figs. 4 and 6). As summarized in Fig. 4, other hydrophobic amino acids were also capable of replacing methionine. As anticipated, replacement of methionine with a charged amino acid (glutamate) failed to function, as did replacement with alanine, whose R-group hydrophobicity is significantly lower than methionine or valine. In addition, replacement with cysteine, the other

sulphur containing amino acid, failed to support SRP function (Fig. 4-Bottom row). None of the methionine replacements significantly altered protein levels (Fig. 5).

To more rigorously assess the ability of the amino acid replacements to support SRP function, we constructed a series of F′ factors to place the *ffh*α4M alleles in single copy so that complementation tests could be performed at physiological levels of expression. While several of the mutants complemented in SLD106 at near wild type levels when expressed from a multiple copy-number plasmid (Fig. 4), use of F′ factors revealed additional differences between the mutants. For example, while *ffh*α4M→V, *ffh*α4M→F, *ffh*α4M→Y and *ffh*α4M→W and *ffh*α4M→I were all able to complement in single copy (Fig. 6-Top row), the isoleucine and tryptophan replacements complemented only at 42^{o}C (Fig. 6-Top row and Table 2). This result is consistent with the observation that amino acids with less flexibility are substituted for methionine residues in the M-domain in organisms that grow at higher temperatures. Conversely, the tyrosine replacement mutant complemented best at 30^{o}C (Fig. 6-Top row and Table 2). Each of the mutants were also shown to have varying effects on SRP-dependent protein targeting (Fig. 7A), as measured by monitoring targeting of the SRP-dependent cytoplasmic membrane protein FtsQ by biotinylation and by measuring β-galactosidase activity of the product of an *ftsQ-lacZ* gene fusion (Fig. 7). These differences correlated with the growth rate of each mutant when the *ffh*α4M alleles were expressed in single copy. Collectively, the results reveal that SRP function in each of the mutants compared to wild type is ordered as follows: *ffh*$^{+}$ (wild type) > *ffh*α4M→V > *ffh*α4M→F > *ffh*α4M→Y > *ffh*α4M→W > *ffh*α4M→I > *ffh*α4M→L > *ffh*α4M→A, *ffh*α4M→E, *ffh*α4M→C.

Because the extreme carboxy-terminus was shown to be highly variable among all of the microorganisms used in this study, we also determined that this region of Ffh was important for

SRP function. A mutant deleted for the carboxy-terminal tail was viable only when expressed from a muliple copy number plasmid (data not shown).

Since *ffh*α4M→V grew near wild type levels (Fig. 6-Top row and Fig. 7), we further tested its ability to fully replace methionine by constructing a mutant allele where all 20 of the methionine residues in the M-domain were replaced with valine (Fig. 1D). This "valine thistle" mutant failed to support cell viability in *E. coli*, however. Multiple sequence alignments of the M-domain revealed three amino acids, corresponding to positions 344, 383 and 426 of *E. coli* Ffh, were highly conserved. Interestingly, position 383, found in helix αM3 was invariant in all four groups. We found that restoring methionine at these three positions in *ffh*M→Vα1-4, yielding allele *ffh*V→M(x3), expressed a protein that was able to complement *ffh*::*kan1* in SLD106 when expressed in both multiple and single copy (Fig. 6-Bottom row).

The results presented were not anticipated based upon phylogenetic sequence analysis. Although valine is structurally similar to leucine and isoleucine, its R group is shorter and comparatively less hydrophobic, it possesses sufficient properties to support SRP function at nearly wild type levels. While some proteins may be more dependent upon methionine bristles for efficient membrane localization, polypeptides essential for bacterial growth must be less dependent for efficient targeting. While methionine side chains may possess the balance of side chain length and hydrophobicity for efficient signal peptide recognition, as suggested by its high degree of conservation, SRP can function with several hydrophobic R-groups in the M-domain. The failure of leucine and isoleucine, at lower growth temperatures, to replace methionine may reflect the loss of helical structure in the M-domain and may not be directly related to the inability of these amino acids to bind signal peptides. Biochemical experiments will be required to test this. Also, it is possible that of the three most highly conserved methionines in the M-

domain (Fig. 1D), not all of them may be required for Ffh function. Of these, methionine-383 does not appear to contact signal peptides (Fig. 1A) and may be important for maintaining the structure of the highly packed helical region of the M-domain.

ACKNOWLEDGEMENTS

Funding for this research was from the National Institutes of Health grant R01 GM069628. We thank Gaya Amarasinghe for helpful discussions.

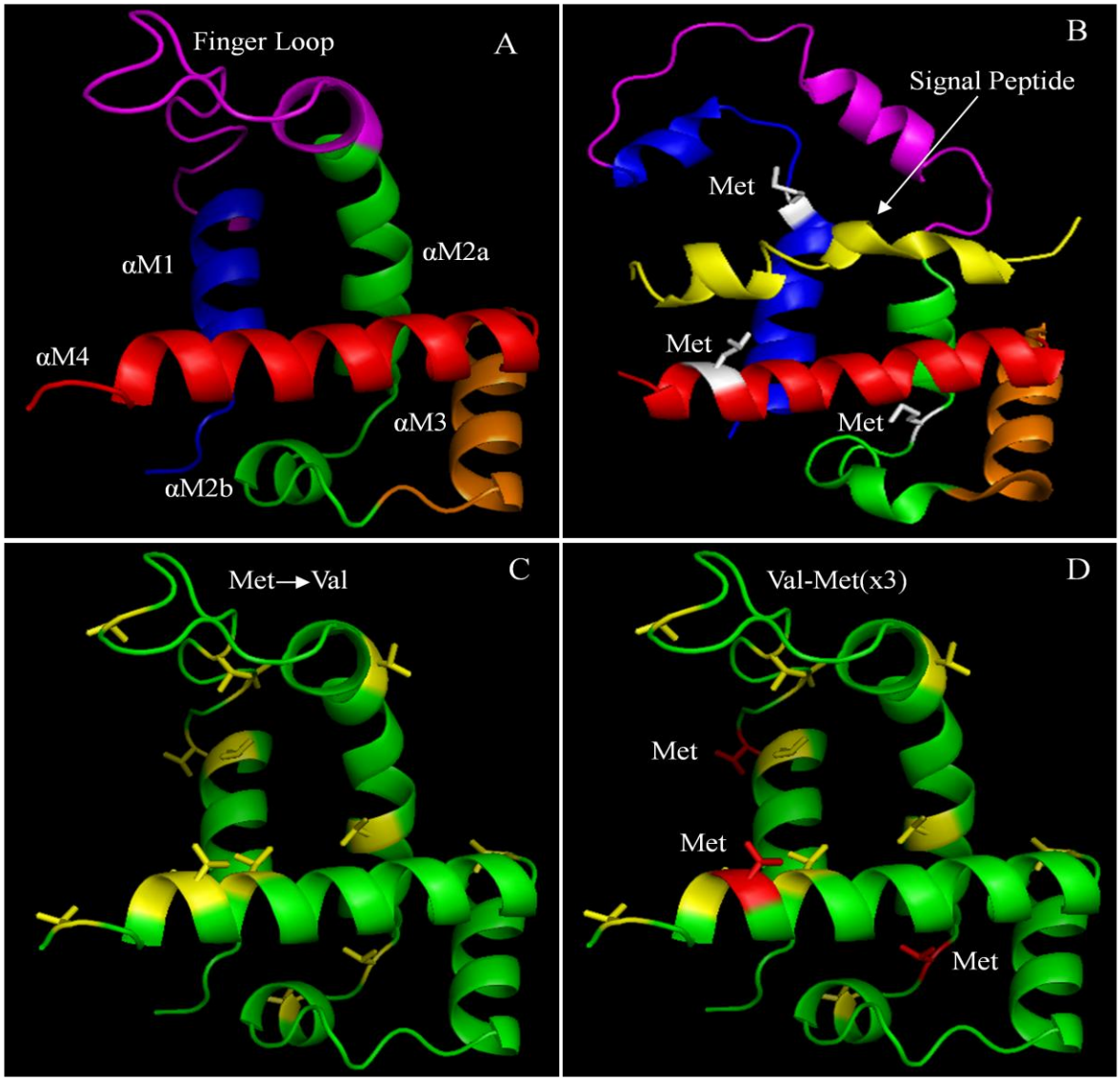

Fig. 1. Structures of the Ffh M-domain. **(A)** Structure of the M-domain of Ffh from *Thermus aquaticus* (Protein Data Bank 2Ffh) showing the signal sequence binding groove formed by αM1 (blue), αM2 (green), αM4 (red) and the finger loop (magenta). Helix αM3 (orange) has been shown to bind the SRP RNA. **(B)** Modified structure of the *Sulfolobus solfataricus* SRP54–signal peptide fusion protein (Protein Data Bank 3KL4). The three highly conserved methionine residues revealed by sequence analysis are shown in white and the signal sequence in shown in yellow. **(C)** Mutant *ffh*M→Vα1-4 modeled using the structure of the M-domain of Ffh from *T. aquaticus* (Protein Data Bank 2Ffh) showing methionine residues that were substituted with valine (yellow). **(D)** Mutant *ffh*V→M(x3) modeled using the structure of the M-domain of Ffh from *T. aquaticus* (Protein Data Bank 2Ffh). The three highly conserved methionine residues are shown in red and valine residues are shown in yellow.

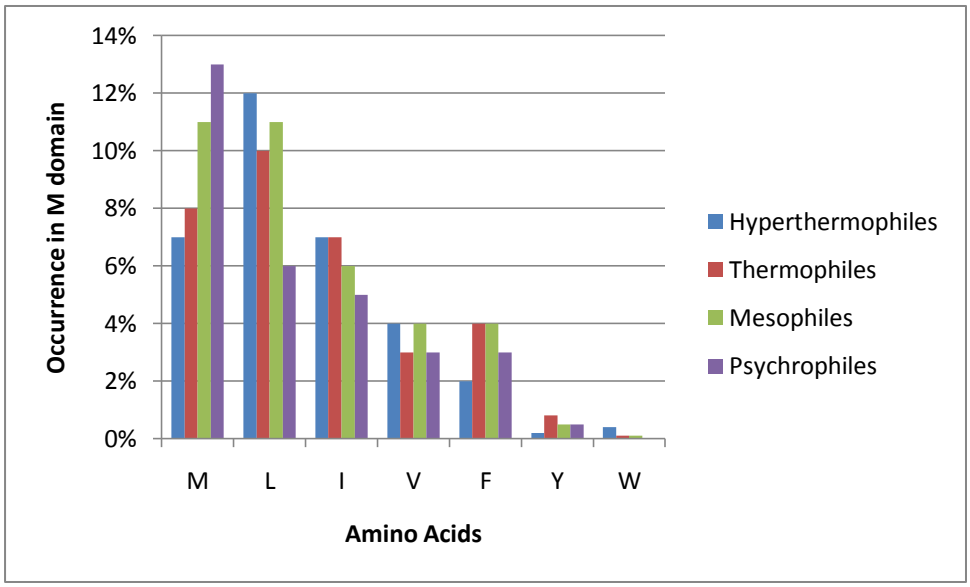

Fig. 2. Representatives of amino acids found in the M-domain of Ffh. The occurrence, shown as percentages, of hydrophobic residues located in the entire M domain of Ffh from hyperthermophiles, thermophiles, mesophiles and psychrophiles.

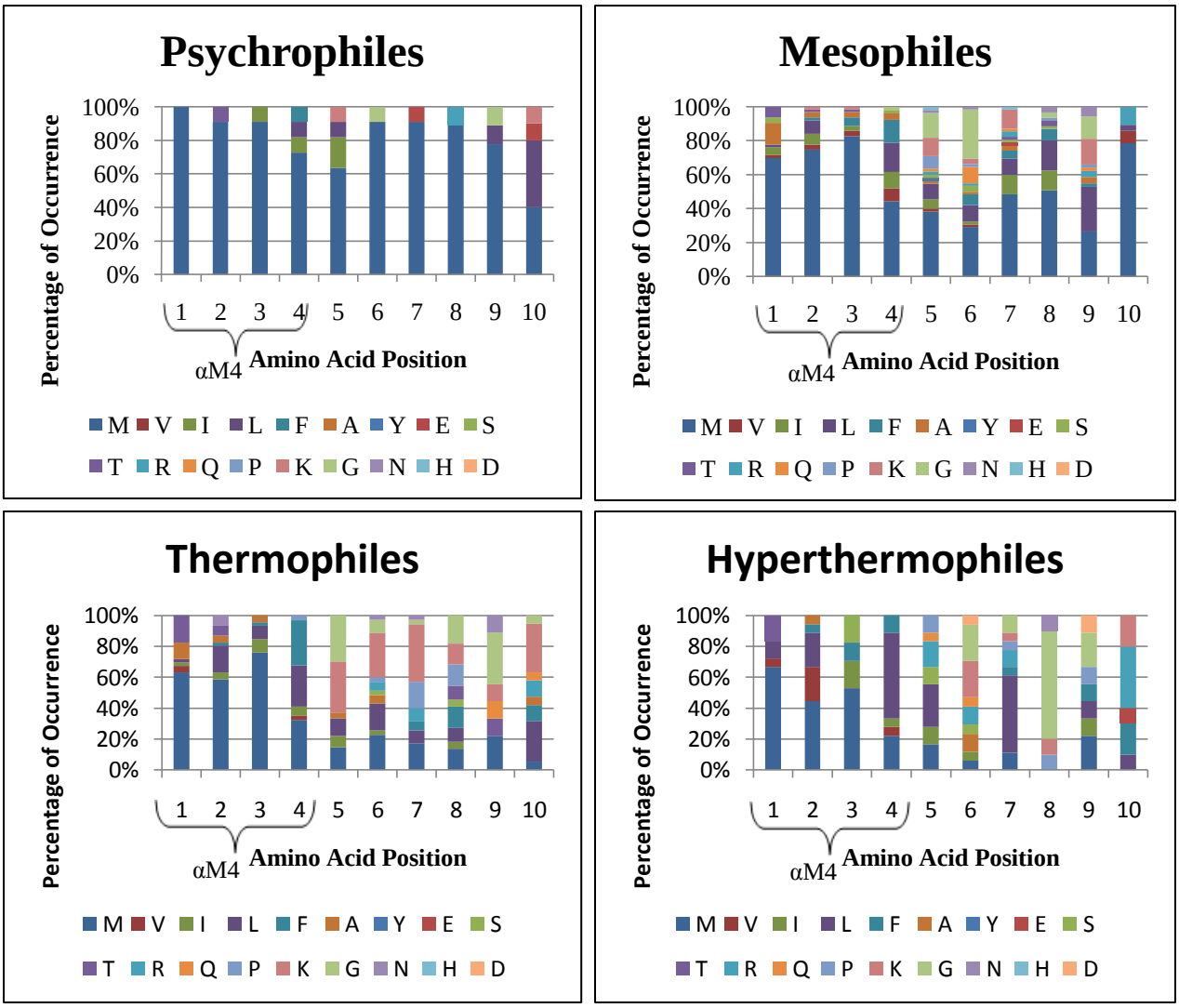

Fig. 3. Comparison of Ffh M-domain sequences from bacteria and archaea representing varying optimal growth temperatures. Four methionine residues in the αM4 domain (positions 423, 426, 427, 430) as well as 6 additional residues at positions 435, 438, 439, 442, 445 and 446 at the extreme C-terminus from Ffh in *E. coli* were compared to sequences from four groups based upon OGT and using multiple sequence alignments. The percentage of occurrence of each amino acid listed in the legend is represented. Methionine residues among all four groups are found to occur frequently in positions 1, 2, 3 and 4 (found in αM4).

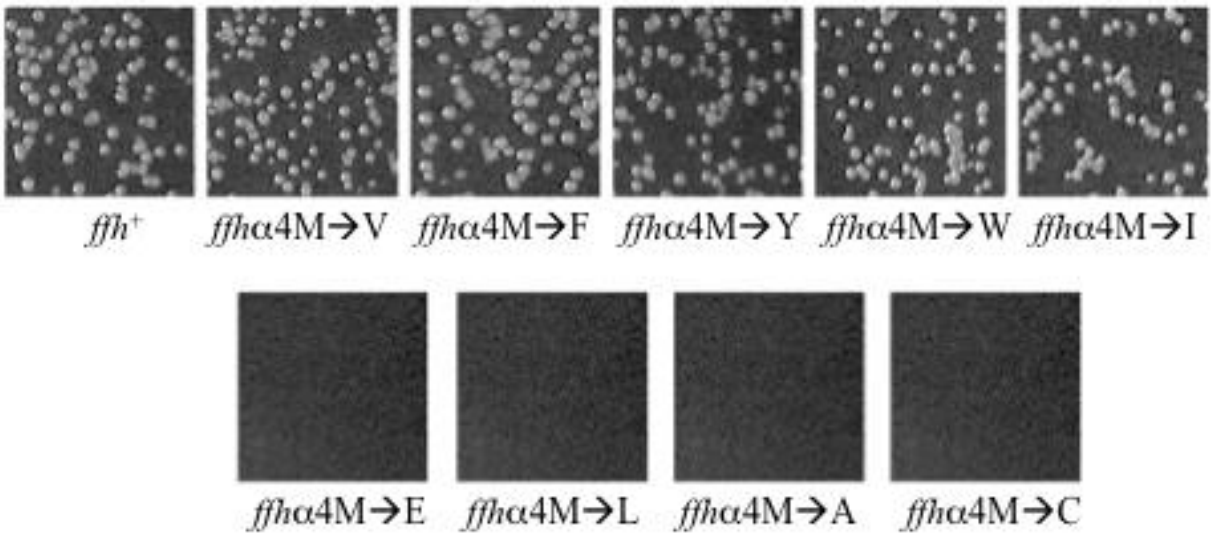

Fig. 4. Growth of *ffh*α4M mutants. Strain SLD106 was transformed with plasmids expressing *ffh*α4M→F, *ffh*α4M→W, *ffh*α4M→I, *ffh*α4M→Y, *ffh*α4M→V and the positive control (*ffh*$^+$), as shown, in multiple copy numbers and incubated at 42°C (Top Row). Mutant alleles *ffh*α4M→L, *ffh*α4M→A, *ffh*α4M→C, and *ffh*α4M→E failed to complement (Bottom Row).

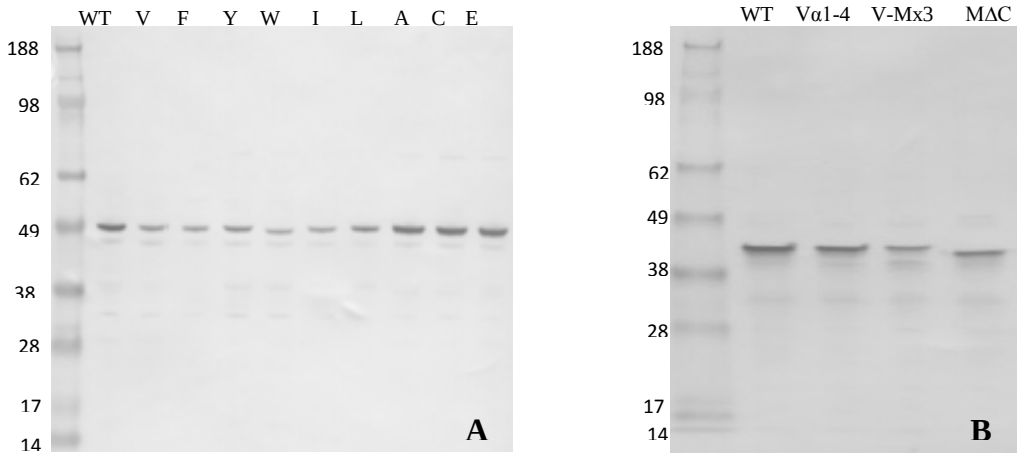

Fig. 5. Detection of products of mutant *ffh* alleles. Western blot analysis was used to detect the expression of wild type *ffh* and *ffh*α4M mutants **(A)** *ffh*$^{+}$ (WT), *ffh*α4M→V (V), *ffh*α4M→F (F), *ffh*α4M→Y (Y), *ffh*α4M→W (W), *ffh*α4M→I (I), *ffh*α4M→L (L), *ffh*α4M→A (A), *ffh*α4M→C (C) and *ffh*α4M→E (E) and **(B)** *ffh*$^{+}$ (WT), *ffh*M→Vα1-4 (Vα1-4), *ffh*V→M(x3) (V-Mx3) and *ffh*MΔC (MΔC) using antibody against the cMyc epitope. Differences in protein size in (A) and (B) are a result of running the protein samples on different types of protein gels. The same molecular weight marker was used in (A) and (B).

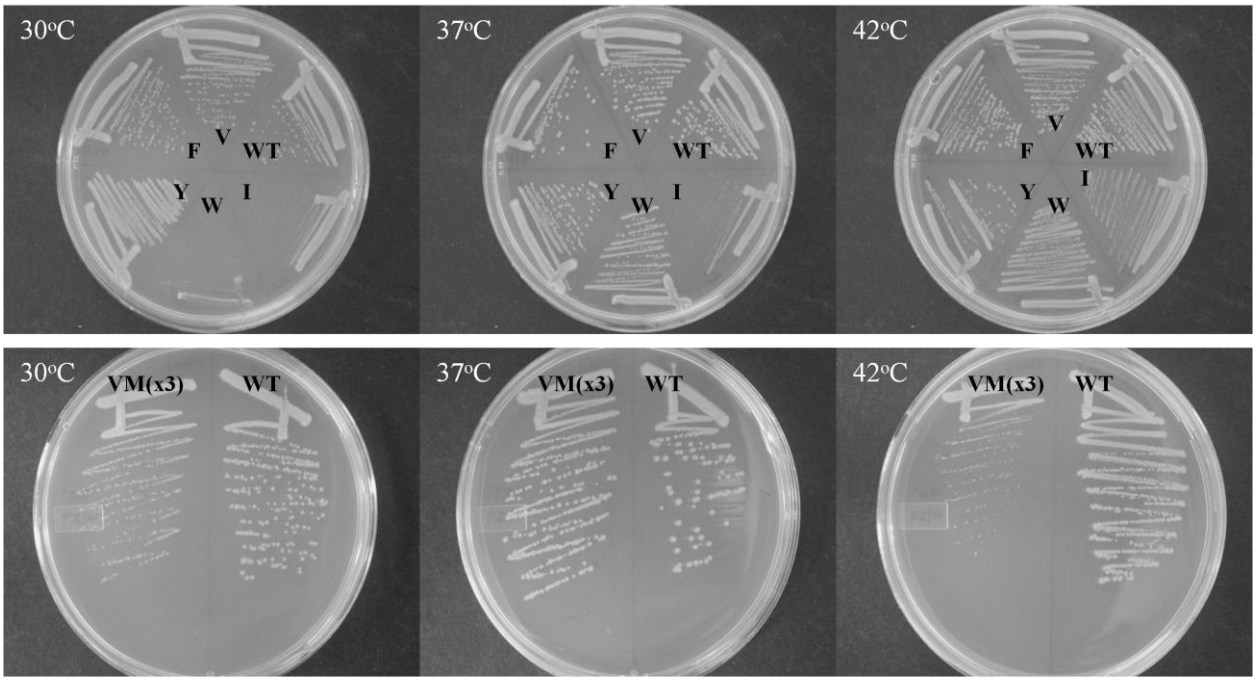

Fig. 6. Phenotypes of *ffh* M-domain mutants. (**Top Row**) Mutants representing αM4-C terminus mutants in single copy. Colonies grown and restreaked at 30°C, 37°C and 42°C showing positive control WT (*ffh*$^{+}$α4), F (*ffh*α4M→F), W (*ffh*α4M→W), I (*ffh*α4M→I), Y (*ffh*α4M→Y) and V (*ffh*α4M→V). (**Bottom Row**) Mutant VM(x3) (*ffh*V→M(x3)) expressed in single copy. Colonies grown and restreaked at 30°C, 37°C and 42°C showing positive control (*ffh*$^{+}$α4) and *ffh*V→M(x3).

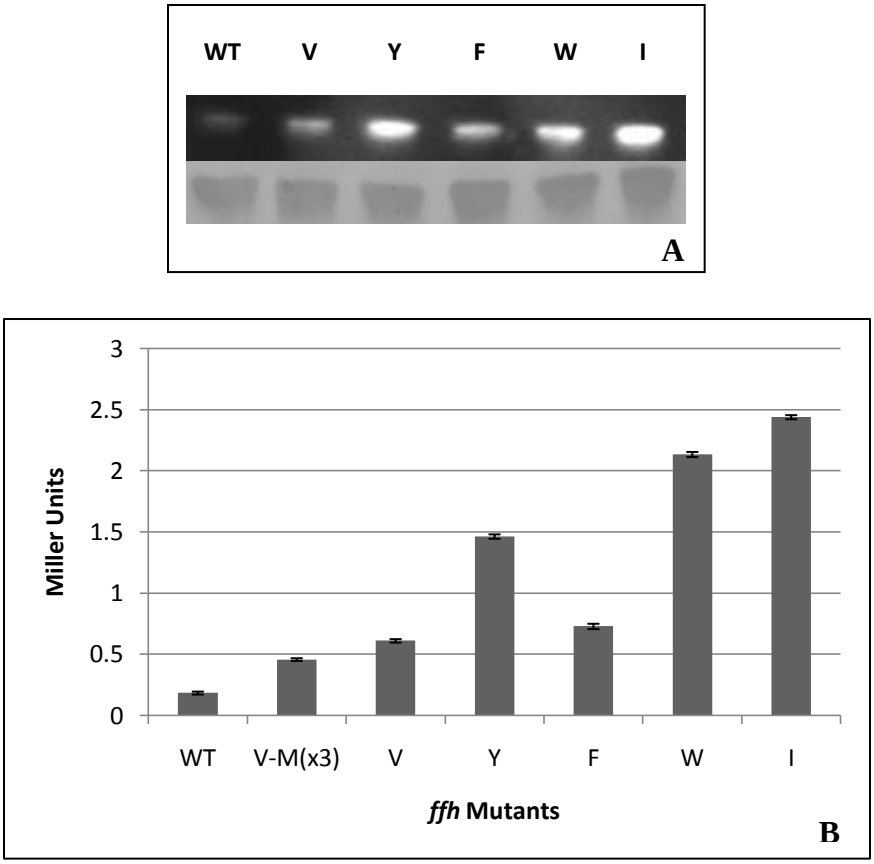

Fig 7. SRP activity of *ffh* mutants. (**A**) Levels of biotinylation of FtsQ, WT (wild type), V (*ffh*α4M→V), Y (*ffh*α4M→Y), F (*ffh*α4M→F), W (*ffh*α4M→W), I (*ffh*α4M→I). Wild type and mutants were tested for the amount of biotinylated inner membrane protein FtsQ. (Top panel) Biotinylated FtsQ detected using streptavidin-horseradish peroxidase conjugate. (Bottom panel) FtsQ detected by Western Blot analysis using Rabbit-anti-V5 primary antibody. (**B**) β-galactosidase activity of the *ffh* M domain mutants was measured and reported in Miller units.

Table 1: Strains and plasmids used in this study

Strain of plasmid		Relevant genotype or description	Source or Reference
E. coli strain			
	NEB5α	*fhuA2*Δ(*argF-lacZ*)*U169 phoA glnV44 80* Δ(*lacZ*)*M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17* (general cloning host)	New England Biolabs
	MC4100	F$^-$ *araD139*, Δ(*argF-lac*)U169, *rspL150*, *relA1*, *flbB5301*, *fruA25*, *deoC1*, *ptsF25* e14-	Lab collection
	WAM100	MC4100, *ara*$^+$	(29)
	WAM121	WAM100, *attB*:: P_{araBAD}-*ffh*$^+$, *ffh*::*kan*1 (source of *ffh*::*kan*1 allele)	(9)
	SLD106	WAM100, *ffh*::*kan*1, p*ffh*TS-Spc (Spc^R) (Tet^R)	This study
	ECF529	Δ *araBAD*, Δ*rhaBAD*, Δ*araFGH*, Δ*araE*, *rrnB*PI(CTC-AGA)-*lacYA177C*	(5)
	XLU102	ECF529, *bla*:: Δ*kan*	Lab collection
	SLD108	XLU102, *ffh*::*kan*1, p*ffh*TS-Spc	This study
	CSH100	F' *lac proA*$^+$*B*$^+$ (*lacI*^{q} *lacPL8*)/*araD (gpt-lac)5* (source of F' *lac*)	(23)
Plasmid			
	p*ffh*TS-Spc	pSC101ts, *ffh*$^+$*spc* (Spc^r)	Lab collection
	pSLD-*ffh*10	*ffh*$^+$, *bla* (Amp^R), ColE1	This study
	pSLD-*ffh*11	*ffh*α4M→A, Amp^R, ColE1	This study
	pSLD-*ffh*12	*ffh*α4M→C, Amp^R, ColE1	This study
	pSLD-*ffh*13	*ffh*α4M→E, Amp^R, ColE1	This study
	pSLD-*ffh*14	*ffh*α4M→I, Amp^R, ColE1	This study
	pSLD-*ffh*15	*ffh*α4M→L, Amp^R, ColE1	This study

Table 1: (continued)

Strain of plasmid		Relevant genotype or description	Source or Reference
Plasmid			
	pSLD-*ffh*16	*ffh*α4M→F, Amp^R, ColE1	This study
	pSLD-*ffh*17	*ffh*α4M→W, Amp^R, ColE1	This study
	pSLD-*ffh*18	*ffh*α4M→Y, Amp^R, ColE1	This study
	pSLD-*ffh*19	*ffh*α4M→V, Amp^R, ColE1	This study
	pSLD-*ffh*20	*ffh*MΔC(amber mutation at position 436), Amp^R, ColE1	This study
	pSLD-*ffh*21	*ffh*M→Vα1-4, Amp^R, ColE1	This study
	pSLD-*ffh*22	*ffh*V→M(x3)(positions 344, 383 and 426), Amp^R, ColE1	This study
	pLac-*ffh*	Vector for expression of *ffh* under P_{lac} control,Amp^R	Lab collection
	pBAD*ftsQ*-V5-PSBT		(28)

Table 2: Ffh α4M domain mutants

***ffh* Allele Designation**	**Amino Acid Sequence**	**GRC 30°C**	**GRC 37°C**	**GRC 42°C**
ffh^{+}α4	DD**M**QR**MM**KK**M**KKGG**M**AK**MM**RS**M**KG**MM**PPGFPGR	0.75	0.97	1.04
*ffh*α4M→A	DD**A**QR**AA**KK**A**KKGG**A**AK**AA**RS**A**KG**AA**PPGFPGR	---	---	---
*ffh*α4M→C	DD**C**QR**CC**KK**C**KKGG**C**AK**CC**RS**C**KG**CC**PPGFPGR	---	---	---
*ffh*α4M→E	DD**E**QR**EE**KK**E**KKGG**E**AK**EE**RS**E**KG**EE**PPGFPGR	---	---	---
*ffh*α4M→I	DD**I**QR**II**KK**I**KKGG**I**AK**II**RS**I**KG**II**PPGFPGR	---	---	0.66
*ffh*α4M→L	DD**L**QR**LL**KK**L**KKGG**L**AK**LL**RS**L**KG**LL**PPGFPGR	---	---	---
*ffh*α4M→F	DD**F**QR**FF**KK**F**KKGG**F**AK**FF**RS**F**KG**FF**PPGFPGR	0.71	0.98	1.04
*ffh*α4M→Y	DD**Y**QR**YY**KK**Y**KKGG**Y**AK**YY**RS**Y**KG**YY**PPGFPGR	0.73	0.96	0.42
*ffh*α4M→W	DD**W**QR**WW**KK**W**KKGG**W**AK**WW**RS**W**KG**WW**PPGFPGR	0.34	0.92	1.00
*ffh*α4M→V	DD**V**QR**VV**KK**V**KKGG**V**AK**VV**RS**V**KG**VV**PPGFPGR	0.73	0.97	0.96

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389-402.
2. **Batey, R. T., R. P. Rambo, L. Lucast, B. Rha, and J. A. Doudna.** 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science **287:**1232-9.
3. **Bernstein, H. D., M. A. Poritz, K. Strub, P. J. Hoben, S. Brenner, and P. Walter.** 1989. Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature **340:**482-6.
4. **Bowers, C. W., F. Lau, and T. J. Silhavy.** 2003. Secretion of LamB-LacZ by the signal recognition particle pathway of *Escherichia coli*. J. Bacteriol. **185:**5697-705.
5. **Bowers, L. M., K. Lapoint, L. Anthony, A. Pluciennik, and M. Filutowicz.** 2004. Bacterial expression system with tightly regulated gene expression and plasmid copy number. Gene **340:**11-8.
6. **Brown, S., and M. J. Fournier.** 1984. The 4.5S RNA gene of *Escherichia coli* is essential for cell growth. J. Mol. Biol. **178:**533-50.
7. **Clemons, W. M., Jr., K. Gowda, S. D. Black, C. Zwieb, and V. Ramakrishnan.** 1999. Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 Å resolution: evidence for the mechanism of signal peptide binding. J. Mol. Biol. **292:**697-705.
8. **Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner.** 2004. WebLogo: a sequence logo generator. Genome Res. **14:**1188-90.
9. **de Gier, J. W., P. Mansournia, Q. A. Valent, G. J. Phillips, J. Luirink, and G. von Heijne.** 1996. Assembly of a cytoplasmic membrane protein in *Escherichia coli* is dependent on the signal recognition particle. FEBS Lett. **399:**307-9.
10. **Doudna, J. A., and R. T. Batey.** 2004. Structural insights into the signal recognition particle. Annu. Rev. Biochem. **73:**539-57.
11. **Driessen, A. J., and N. Nouwen.** 2008. Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. **77:**643-67.
12. **Egea, P. F., R. M. Stroud, and P. Walter.** 2005. Targeting proteins to membranes: structure of the signal recognition particle. Curr. Opin. Struct. Biol. **15:**213-20.
13. **Freymann, D. M., R. J. Keenan, R. M. Stroud, and P. Walter.** 1997. Structure of the conserved GTPase domain of the signal recognition particle. Nature **385:**361-4.
14. **Ilangovan, U., S. H. Bhuiyan, C. S. Hinck, J. T. Hoyle, O. N. Pakhomova, C. Zwieb, and A. P. Hinck.** 2008. *A. fulgidus* SRP54 M-domain. J. Biomol. NMR **41:**241-8.
15. **Janda, C. Y., J. Li, C. Oubridge, H. Hernandez, C. V. Robinson, and K. Nagai.** 2010. Recognition of a signal peptide by the signal recognition particle. Nature.
16. **Jander, G., J. E. Cronan, Jr., and J. Beckwith.** 1996. Biotinylation *in vivo* as a sensitive indicator of protein secretion and membrane protein insertion. J. Bacteriol. **178:**3049-58.
17. **Keenan, R. J., D. M. Freymann, R. M. Stroud, and P. Walter.** 2001. The signal recognition particle. Annu. Rev. Biochem. **70:**755-75.

18. **Keenan, R. J., D. M. Freymann, P. Walter, and R. M. Stroud.** 1998. Crystal structure of the signal sequence binding subunit of the signal recognition particle. Cell **94:**181-91.
19. **Luirink, J., C. M. ten Hagen-Jongman, C. C. van der Weijden, B. Oudega, S. High, B. Dobberstein, and R. Kusters.** 1994. An alternative protein targeting pathway in *Escherichia coli*: studies on the role of FtsY. Embo. J. **13:**2289-96.
20. **Lutcke, H., S. High, K. Romisch, A. J. Ashford, and B. Dobberstein.** 1992. The methionine-rich domain of the 54 kDa subunit of signal recognition particle is sufficient for the interaction with signal sequences. Embo. J. **11:**1543-51.
21. **Miller, J. D., H. D. Bernstein, and P. Walter.** 1994. Interaction of *E. coli* Ffh/4.5S ribonucleoprotein and FtsY mimics that of mammalian signal recognition particle and its receptor. Nature **367:**657-9.
22. **Miller, J. D., H. Wilhelm, L. Gierasch, R. Gilmore, and P. Walter.** 1993. GTP binding and hydrolysis by the signal recognition particle during initiation of protein translocation. Nature **366:**351-4.
23. **Miller, J. H.** 1992. A short course in bacterial genetics: a laboratory manual and handbook for *Escherichia coli* and related bacteria, vol.1. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
24. **Montoya, G., C. Svensson, J. Luirink, and I. Sinning.** 1997. Crystal structure of the NG domain from the signal-recognition particle receptor FtsY. Nature **385:**365-8.
25. **Oh, D. B., G. S. Yi, S. W. Chi, and H. Kim.** 1996. Structure of a methionine-rich segment of *Escherichia coli* Ffh protein. FEBS Lett. **395:**160-4.
26. **Park, S. K., F. Jiang, R. E. Dalbey, and G. J. Phillips.** 2002. Functional analysis of the signal recognition particle in *Escherichia coli* by characterization of a temperature-sensitive *ffh* mutant. J. Bacteriol. **184:**2642-53.
27. **Peluso, P., S. O. Shan, S. Nock, D. Herschlag, and P. Walter.** 2001. Role of SRP RNA in the GTPase cycles of Ffh and FtsY. Biochemistry **40:**15224-33.
28. **Peterson, J. M., and G. J. Phillips.** 2008. Characterization of conserved bases in 4.5S RNA of *Escherichia coli* by construction of new F' factors. J. Bacteriol. **190:**7709-18.
29. **Phillips, G. J., and T. J. Silhavy.** 1992. The *E. coli ffh* gene is necessary for viability and efficient protein export. Nature **359:**744-6.
30. **Poritz, M. A., H. D. Bernstein, K. Strub, D. Zopf, H. Wilhelm, and P. Walter.** 1990. An *E. coli* ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. Science **250:**1111-7.
31. **Powers, T., and P. Walter.** 1995. Reciprocal stimulation of GTP hydrolysis by two directly interacting GTPases. Science **269:**1422-4.
32. **Ribes, V., K. Romisch, A. Giner, B. Dobberstein, and D. Tollervey.** 1990. *E. coli* 4.5S RNA is part of a ribonucleoprotein particle that has properties related to signal recognition particle. Cell **63:**591-600.
33. **Romisch, K., J. Webb, J. Herz, S. Prehn, R. Frank, M. Vingron, and B. Dobberstein.** 1989. Homology of 54K protein of signal-recognition particle, docking protein and two *E. coli* proteins with putative GTP-binding domains. Nature **340:**478-82.
34. **Romisch, K., J. Webb, K. Lingelbach, H. Gausepohl, and B. Dobberstein.** 1990. The 54-kD protein of signal recognition particle contains a methionine-rich RNA binding domain. J. Cell. Biol. **111:**1793-802.

35. **Rosendal, K. R., K. Wild, G. Montoya, and I. Sinning.** 2003. Crystal structure of the complete core of archaeal signal recognition particle and implications for interdomain communication. Proc. Natl. Acad. Sci. USA **100:**14701-6.
36. **Shan, S. O., R. M. Stroud, and P. Walter.** 2004. Mechanism of association and reciprocal activation of two GTPases. PLoS Biol. **2:**e320.
37. **Tian, H., and J. Beckwith.** 2002. Genetic screen yields mutations in genes encoding all known components of the *Escherichia coli* signal recognition particle pathway. J. Bacteriol. **184:**111-8.
38. **Tian, H., D. Boyd, and J. Beckwith.** 2000. A mutant hunt for defects in membrane protein assembly yields mutations affecting the bacterial signal recognition particle and Sec machinery. Proc. Natl. Acad. Sci. USA **97:**4730-5.
39. **Walter, P., and G. Blobel.** 1980. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. Proc. Natl. Acad. Sci. USA **77:**7112-6.
40. **Walter, P., and G. Blobel.** 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. Nature **299:**691-8.
41. **Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton.** 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics **25:**1189-91.
42. **Wild, K., M. Halic, I. Sinning, and R. Beckmann.** 2004. SRP meets the ribosome. Nat. Struct. Mol. Biol. **11:**1049-53.
43. **Zopf, D., H. D. Bernstein, A. E. Johnson, and P. Walter.** 1990. The methionine-rich domain of the 54 kd protein subunit of the signal recognition particle contains an RNA binding site and can be crosslinked to a signal sequence. Embo. J. **9:**4511-7.

CHAPTER 4. Introduction

In the mid 1970's two independent methods of DNA sequencing were introduced, including the Sanger sequencing method, named after its developer Frederick Sanger (16, 17), and the chemical sequencing method developed by Maxam and Gilbert (13). These technologies have had a major impact on biological research ever since. In particular, Sanger sequencing became the prominent method for DNA sequence production for over 30 years, including sequencing of the human genome.

In the past five years new sequencing technologies, referred to as "next-generation sequencing", have emerged and are revolutionizing biological research. Several next generation sequencing platforms are commercially available including Roche/454 Genome Sequencer (9), Illumina/Solexa Genome Analyzer II (4, 22), Applied Biosystems SOLiD System (20). These massively parallel sequencing platforms allow for rapid and cost effective generation of sequencing data. Owing to the significant reduction in cost, next generation sequencing has provided a way for individual laboratories to carry out projects to address questions on a genome scale, including metagenomics, ancient DNA research, transcriptome analysis, and mapping of DNA-protein interactions, that previously could only be pursued by genomic centers (7, 8, 18, 19).

Next generation sequencing technology has had a significant impact on bacterial genomics. Several bacterial genome sequencing projects of multiple isolates of the same species have revealed extensive intraspecies genotypic heterogeneity (1, 5, 6, 21), which has been noted to have the potential to impact vaccine development and discovery of novel antimicrobials (14).

Despite the value of using next generation sequencing technology, the amount of genome sequence data and the rate of data generation presents bioinformatic challenges such as developing tools to organize, store, manage and analyze annotated sequencing data. In the past few years, several tools have been developed to address this need including Integrated Microbial Genomes (IMG) system (10, 12), Integrated Microbial Genomes-Expert Review (IMG ER) system (11), the Microbial Genome Database (MBGD) (23, 24), the Comprehensive Microbial Resource (CMR) (15), and the EDGAR software (2), just to name a few.

Although the previously mentioned software tools offer sophisticated sequence data analysis functionalities, they require researchers to develop a working knowledge of the software functionalities or to have access to the computational support needed to utilize the software. However, due to the increase in availability and use of next-generation sequencing technologies in academic, industrial and government settings computational support for individual labs may not be readily available. To address some of these limitations with currently available computational tools, we developed the **Dra**ft **Ge**n**o**me **E**valuation **T**ool (DraGnET). The following chapter details the development and use of the DraGnET software (3) for storage, management and preliminary comparative analysis of bacterial genome sequences.

REFERENCES

1. **Bjorkholm, B., A. Lundin, A. Sillen, K. Guillemin, N. Salama, C. Rubio, J. I. Gordon, P. Falk, and L. Engstrand.** 2001. Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. Infect. Immun. **69:**7832-8.
2. **Blom, J., S. P. Albaum, D. Doppmeier, A. Puhler, F. J. Vorholter, M. Zakrzewski, and A. Goesmann.** 2009. EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics **10:**154.
3. **Duncan, S., R. Sirkanungo, L. Miller, and G. J. Phillips.** 2010. DraGnET: software for storing, managing and analyzing annotated draft genome sequence data. BMC Bioinformatics **11:**100.
4. **Fedurco, M., A. Romieu, S. Williams, I. Lawrence, and G. Turcatti.** 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. **34:**e22.
5. **Fitzgerald, J. R., D. E. Sturdevant, S. M. Mackie, S. R. Gill, and J. M. Musser.** 2001. Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. Proc. Natl. Acad. Sci. USA **98:**8821-6.
6. **Fukiya, S., H. Mizoguchi, T. Tobe, and H. Mori.** 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic hybridization microarray. J .Bacteriol. **186:**3911-21.
7. **Mardis, E. R.** 2008. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. **9:**387-402.
8. **Marguerat, S., B. T. Wilhelm, and J. Bahler.** 2008. Next-generation sequencing: applications beyond genomes. Biochem. Soc. Trans. **36:**1091-6.
9. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376-80.
10. **Markowitz, V. M., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, and N. C. Kyrpides.** 2006. The integrated microbial genomes (IMG) system. Nucleic Acids Res. **34:**D344-8.
11. **Markowitz, V. M., K. Mavromatis, N. N. Ivanova, I. M. Chen, K. Chu, and N. C. Kyrpides.** 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics **25:**2271-8.
12. **Markowitz, V. M., E. Szeto, K. Palaniappan, Y. Grechkin, K. Chu, I. M. Chen, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, N. N. Ivanova, and N. C.**

Kyrpides. 2008. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res. **36:**D528-33.

13. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74:**560-4.
14. **Muzzi, A., V. Masignani, and R. Rappuoli.** 2007. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. Drug Discov. Today **12:**429-39.
15. **Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res. **29:**123-5.
16. **Sanger, F., and A. R. Coulson.** 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. **94:**441-8.
17. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463-7.
18. **Schuster, S. C.** 2008. Next-generation sequencing transforms today's biology. Nat. Methods **5:**16-8.
19. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing. Nat. Biotechnol **26:**1135-45.
20. **Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church.** 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science **309:**1728-32.
21. **Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser.** 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA **102:**13950-5.
22. **Turcatti, G., A. Romieu, M. Fedurco, and A. P. Tairi.** 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. **36:**e25.
23. **Uchiyama, I.** 2007. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res. **35:**D343-6.
24. **Uchiyama, I.** 2003. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. **31:**58-62.

CHAPTER 5. DraGnET: Software for storing, managing and analyzing annotated draft genome sequence data

A paper published in *BMC Bioinformatics*

Stacy Duncan[1, 3], Ruchita Sirkanungo[2], Leslie Miller[2,3], Gregory J. Phillips[*1,3]

[1] Primary author, planned, wrote and tested the software, Graduate student and Professor, respectively, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, Iowa, USA

[2] Secondary author, planned, wrote and tested the software, Graduate student and Professor, respectively, Department of Computer Science, Iowa State University, Ames, Iowa, USA

[3] Interdepartmental Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA

*Corresponding author

ABSTRACT

Background

New "next generation" DNA sequencing technologies offer individual researchers the ability to rapidly generate large amounts of genome sequence data at dramatically reduced costs. As a result, a need has arisen for new software tools for storage, management and analysis of genome sequence data. Although bioinformatic tools are available for the analysis and management of genome sequences, limitations still remain. For example, restrictions on the submission of data and use of these tools may be imposed, thereby making them unsuitable for sequencing projects that need to remain in-house or proprietary during their initial stages. Furthermore, the availability and use of next generation sequencing in industrial, governmental and academic

environments requires biologists to have access to computational support for the curation and analysis of the data generated; however, this type of support is not always immediately available.

Results

To address these limitations, we have developed DraGnET (Draft Genome Evaluation Tool). DraGnET is an open source web application which allows researchers, with no experience in programming and database management, to setup their own in-house projects for storing, retrieving, organizing and managing annotated draft and complete genome sequence data. The software provides a web interface for the use of BLAST, allowing users to perform preliminary comparative analysis among multiple genomes. We demonstrate the utility of DraGnET for performing comparative genomics on closely related bacterial strains. Furthermore, DraGnET can be further developed to incorporate additional tools for more sophisticated analyses.

Conclusions

DraGnET is designed for use either by individual researchers or as a collaborative tool available through Internet (or Intranet) deployment. For genome projects that require genome sequencing data to initially remain proprietary, DraGnET provides the means for researchers to keep their data in-house for analysis using local programs or until it is made publicly available, at which point it may be uploaded to additional analysis software applications. The DraGnET home page is available at http://www.dragnet.cvm.iastate.edu and includes example files for examining the functionalities, a link for downloading the DraGnET setup package and a link to the DraGnET source code hosted with full documentation on SourceForge.

BACKGROUND

DNA sequencing technology using chain-terminating dideoxy nucleoside triphosphates, first developed by Frederick Sanger [1, 2], has remained the mainstay of genome sequencing efforts for more than thirty years. However, recently developed, new massively parallel DNA sequencing platforms are now extensively used to generate sequence data at a fraction of the cost and labor required by Sanger technology. Three "next generation" sequencing systems that are currently commercially available include the Roche/454 Genome Sequencer [3], Illumina/Solexa Genome Analyzer II [4, 5] and Applied Biosystems SOLiD System [6]. In addition, commercial release of two additional platforms, including the Helicos Heliscope and the Pacific Biosmart SMRT, are planned for 2010 [7].

Collectively, these systems, with their high depth of coverage and relatively low costs, have allowed individual researchers to initiate genome sequencing projects that were previously available to only large genome centers [8-10]. The enhanced sequencing capability afforded by next-generation sequencing has had an especially significant impact on bacterial genomics. By facilitating genome sequencing of multiple isolates of the same bacterial species, several examples of extensive intraspecies genotypic heterogeneity have been revealed, leading to a revision of many long-standing views of microbial speciation [11-14]. One of the first such studies revealed significant genetic variability among eight different strains of *Streptococcus agalactiae*, group B *Streptococcus* (GBS) [14]. After performing cross strain comparisons Tettelin *et al*. found a considerable number of genes not shared among the strains. Their discovery led to the proposal of the bacterial "pan-genome", defined as the global gene repertoire of a bacterial species comprised of the core genome (the set of genes shared by all the strains of the same bacterial species), the dispensable genome (the set of genes present in some but not all

of the strains) and the strain specific genes (the set of genes found only in a single strain) [14]. Genome heterogeneity has also been noted for species of *Helicobacter pylori*, *Staphylococcus aureus*, and *Escherichia coli* [13, 15, 16]. As noted by Muzzi *et al.*, comparative genomics of bacterial species has important implications for vaccine development and discovery of novel antimicrobials [17]. Other novel applications for next generation sequencing technologies have also been developed, including bacterial metagenomics [18-20], and transcriptome mapping [21-24].

Despite the potential for new insights into bacterial diversity and function, important challenges continue to include the organization, management and analysis of genome sequencing data. To address the need for tools for querying, analyzing and comparing multiple genomes of related species, several databases and software tools have been developed [25], including the Integrated Microbial Genomes (IMG) system [26, 27], Integrated Microbial Genomes-Expert Review (IMG ER) system [28], GenColors [29, 30], the Microbial Genome Database (MBGD) [31, 32], the Comprehensive Microbial Resource (CMR) [33] and the EDGAR software [34].

The IMG system contains complete and draft microbial genome sequence data generated by the Joint Genomes Institute (JGI) as well as other publicly available genome data not limited to microorganisms. Tools provided through IMG allow users to query, view and perform comparative analysis of genomes, genes and functions. Recently, a new version of IMG called IMG ER has been added to the IMG system. Tools available through IMG ER allow users to analyze and curate annotated microbial genome data whether it is unpublished or published. Although IMG ER allows users to upload their genome sequencing data for curation and analysis, it is not available for download and in-house use. The GenColors software allows users to browse, analyze and compare genome information from complete and ongoing genome

projects related to prokaryotic or eukaryotic genomes. Additionally, GenColors may be used for the purpose of annotation in the case of incomplete projects. The CMR software contains sequence and annotation data for all of the current publicly available completed microbial genomes and provides a variety of comparison tools for the analysis of the multiple genomes including cross-genome analysis capabilities. Currently, however, there is no functionality that allows users to submit genome data for use with CMR. Similar to CMR, MBGD provides users with several tools for the comparison and analysis of complete bacterial genomes. Unlike CMR, MBGD contains a newly added feature called MyMBGD that allows users to add their own genome data to MBGD. The EDGAR software has recently been released and includes comparative analysis tools for the comparison of multiple strains of a given species. EDGAR offers similar capabilities to those found in CMR and MBGD, in addition to features such as phylogenetic analyses and visualization capabilities including Venn diagrams and synteny plots.

While the aforementioned systems include data management and analysis functionalities there are limitations. For example, genome projects that include proprietary data may be restricted in the submission of the data to third party software. Many of the current data management software tools are not available for download and in-house use, a requirement when access to next generation sequencing instruments can outstrip the availability of experienced bioinformaticians to assist with data management and analysis.

In addition to the already mentioned software applications, there are other tools that are designed for genome annotation or re-annotation of unpublished or published genomes [25, 35, 36]. Several of these tools provide data curation capabilities for the purpose of correcting annotation errors and improving annotated data but are restricted to use with the annotated data generated through specified software packages. Additionally, as with many software

applications, they require the researcher to develop a working knowledge of the analysis capabilities of the software as well as provide "expert" curation of the data. With the increased use of next-generation sequencing in academic, industrial and government settings, however, biologists do not always have immediate access to computational support needed to easily manage the data and to initiate comparative analysis.

To overcome some of these limitations, DraGnET was developed specifically to provide biologists with their own web based tool that is both convenient and easy to use. DraGnET allows researchers to independently store, retrieve and curate their own data generated from any annotation engine and to perform genome comparisons during the beginning phase of a sequencing project. Additionally, publicly available genome data can be stored for the purpose of comparing draft genome data with reference genomes. DraGnET includes provisions for data access, searching, and modification as well as access to basic local alignment search tool (BLAST) functionalities [37] for amino acid sequence similarity searches and cross strain comparisons. As a consequence, DraGnET allows investigators to immediately begin testing of biologically relevant hypotheses without having to devote time to learning sophisticated analysis programs or to depend on computational support from designated personnel. Additionally, the DraGnET source code has been made available, allowing researchers to further customize and develop the software to meet the needs of specific sequencing projects.

To demonstrate the utility of DraGnET, we have successfully established a DraGnET project, deployed for Internet access, and performed preliminary cross strain comparisons to identify potential vaccine targets against the animal pathogen *Haemophilus parasuis*. Microbial genome sequencing has proven to be a powerful approach to identify new, protective vaccines via *reverse vaccinology*, i.e., discovery of vaccine targets by scanning sequence data for potential

surface-exposed antigens [38]. Moreover, broadly protective antigens may be identified by comparison of genomes from multiple strains of a single species [17, 39, 40]. Reverse vaccinology has led to the development of new vaccines for several human and animal pathogens where previously vaccines were not available [41-44]. DraGnET enables facile preliminary comparisons of multiple draft or complete genome sequences of any number of organisms, including identification of protein encoding genes shared by multiple strains, making DraGnET a useful bioinformatic tool.

IMPLEMENTATION

The DraGnET web application was developed in Java [see Additional file 1]. DraGnET provides user interfaces for storing information related to strains and their associated annotated gene set in a database. Gene and strain information are stored as objects defined by two Java classes, Gene and Strain (Figure1). The Gene class stores nine gene attributes most of which can be obtained from gene annotation data. The choice of gene attributes was based upon gene information available in public sequence databases such as GenBank and includes additional attributes relevant for vaccine target identification. The Strain class contains information such as the strain name and description. Two additional Java classes, Logininformation class and the Blastdbupdate class are used to define objects related to administrator/curator user information and the date of the last modification made to the data, respectively (Figure 1). Hibernate (version 3.1 core and advanced libraries) is used to map the Java objects, representing the Gene, Strain, Logininformation and Blastdbupdate classes, to relational tables in a MySQL (version 5.0) database. By using Hibernate in the software architecture, DraGnET works with an object database supported by Hibernate. The servlet engine used to support the web interface is

Apache-Tomcat version 6. The web application uses Struts (version 1.2) to implement the Model-View-Controller (MVC) architecture. The MVC architecture provides a way to separate the web interface (view) from the business logic (model) making it easier to implement and modify either component independent of the other. The web interface (view) is implemented through Java Server Pages (JSP). BLAST functionalities are provided by stand-alone executable BLAST software connected to the business logic and web pages are provided for users to interact with BLAST. The BLAST program is configured to run the blastp (protein blast) algorithm and applies the blastall program available from NCBI [45]. The general layout for the architecture of the DraGnET software is provided in Figure 2. The web application was built using MyEclipse version 6.0 and has been successfully tested on Microsoft Windows 2003 and Windows XP operating systems.

DraGnET project setup

A DraGnET project can be installed on a personal computer or it can be setup for Internet (or Intranet) deployment making it a tool that is available for collaborative projects. The initial setup of a DraGnET project requires installation of Java (version 1.6), MySQL (version 5.0) including the MySQL 5.0 GUI Tools, Apache Tomcat 6, and Blast 2.2.18. Executable files for installing all of the aforementioned software are provided in a comprehensive setup package provided through the "DraGnET Application Setup Package" link located on the application's home page (Figure 3). After installing the required software packages the database structure used by Hibernate to map the Java objects to relational tables in the MySQL database is automatically generated by the MySQL 5.0 GUI Tool and a file included in the setup package [see Additional file 2]. This automated process alleviates the requirement of the user to have the

knowledge necessary for setting up the database schema used to store the genome data. After the DraGnET project is set up and genome sequence data has been uploaded into the database, local BLAST databases for each genome need to be formatted for use with the BLAST functionalities provided with the application. Information on all of these steps, as well as additional usage information, is available in the DraGnET_setup.doc provided in the setup package.

RESULTS

DraGnET is an open source web application designed to provide researchers with a tool for storing their own unpublished annotated draft and complete genome data from multiple strains of a species in a database; allowing gene and strain information to be available for retrieving, searching, modifying and downloading. The application also provides a web interface for the use of BLAST, allowing for protein sequence similarity searches and cross strain comparisons of strains stored in the database. In addition, DraGnET provides a link for the automatic generation of FASTA files for each genome stored in the database. The files are available for download and can be used with other software and tools for further analysis. The details of the functionalities of DraGnET are provided in the following section.

Data Management

DraGnET is set up to allow any user to search, view and compare genome sequence data stored in the database; however, only curators may insert and modify the data by signing in to the application. This was designed to prevent inconsistencies in the data and to protect the application when it is being accessed by multiple users from different locations.

Data insertion

Two web pages are provided for the insertion of strain and gene information. The data entry tables for these pages are shown in Figure 4. In the first table the curator enters the strain

information (Figure 4A) and in the second table the curator is directed to upload a file containing gene information for genes contained in the strain (Figure 4B). The application accepts a semicolon-separated plain text file, containing values for the nine gene attributes defined in the Gene class, for batch insertion of gene information into the database. The software then stores the data in the database allowing for subsequent retrievals and updates to be performed.

Data modification

The DraGnET application provides web pages for assigned curators to modify genome data as well as administrator/curator user information. As shown in Figure 5, modifications that can be made to gene and strain data include adding, deleting and updating gene or strain information. The addition and deletion of single or multiple genes to strain(s) already stored in the database follows the same procedure as the addition of a strain and its associated gene set. To delete a single strain the user selects the strain to be deleted and once submitted, the strain information and all of the genes not associated with any other strain are deleted. An important part of data management is the ability to update or modify the information stored in the database, as is the case for draft genome sequences as progress is made toward gap closure and genome completion. To update gene information, the curator enters the gene Id of the gene to be updated (Figure 6A). Subsequently, the gene attributes that need to be modified are selected (Figure 6B). Once the selections are submitted the gene information currently stored in the database is displayed as "old" information and the “new” information may be entered (Figure 6C). A similar procedure is provided for updating strain information.

Formatting BLAST database files

BLAST functionalities for sequence similarity searches and cross strain comparisons are provided through DraGnET web-interfaces. To use these functionalities, BLAST database files

for each strain stored in the database must be created through command line arguments. The command used to format BLAST database files is the similar for each strain stored in the database, having to change only the FASTA file used for BLAST database file generation. Details of this process are included in the DraGnET setup package. Once the BLAST databases are created, all BLAST functionalities offered with DraGnET are available for use.

Data Exploration

The following functionalities are implemented through the web interface and are available for all users.

Quick and Advanced Search

The "Quick Search" option provides users with four different search options for retrieving gene and strain information stored in the database. Searches can be performed by selecting and entering a gene Id, gene name, protein sequence, or strain Id (Figure 7A-B). When a search is performed using a gene Id, gene name or protein sequence the results are displayed in a table containing information for the chosen gene, including the strain that contains the gene (Figure 7C). Searches based upon a strain Id provide the user with strain information as well as the option to download a text file containing gene information for all of the genes contained in the chosen strain. The "Advanced Search" option allows users to search for gene information using more stringent parameters. Users can specify single or multiple gene attributes to use in the advanced search (Figure 8A). Once the attribute(s) are chosen, the user enters search criteria for each attribute chosen and selects the strain(s) they want to search (Figure 8B). If more than one strain is chosen, then the program searches for genes having the same gene identifier and chosen

attributes in common with the set of strains. Search results are written to a text file that can be opened for immediate viewing or saved for future inspection.

BLAST Search

"BLAST Search" provides users with an interface for using the protein BLAST (blastp) program for comparing protein sequences against protein sequence BLAST databases. Each strain stored in the database is used to format BLAST protein databases during the initial setup of a DraGnET project. Subsequently, the strains appear on the BLAST Search page listed under "Search Databases Containing Strains" where users have the option to select single or multiple strain databases to search against (Figure 9A). All parameters are set to NCBI defined default values; however users have the option to refine their search by changing the expectation value (E-value). Users can then input their FASTA formatted query sequence by pasting it into the query box. The output generated will include the input query sequence, the user chosen BLAST database(s) and a list of alignments between the input query sequence and the database hits. The output file is available for immediate viewing and downloading (Figure 9B).

Batch BLAST Search

"Batch BLAST Search" is an extension of "BLAST Search" that allows text files containing multiple FASTA formatted protein sequences (including entire strains) to be uploaded for comparison against a single strain BLAST database (Figure 10). Similar to "BLAST Search", the results are written to a text file available for viewing and downloading.

Batch BLAST Dissimilarity Search

"Batch BLAST Dissimilarity Search" takes as input the results of "Batch BLAST Search" and extracts the gene identifiers associated with protein sequences that produced a "no hits found" result. The resulting set of genes identifiers represents genes that have no protein sequence

homology to any sequences found in the selected search database. Results are written to a text file.

Generate FASTA Files

The "Generate FASTA Files" option automatically generates FASTA files for each strain stored in the database. When users click on the "Generate FASTA Files" button a set of files in FASTA format, one for each strain, will be available to download. Subsequently, the files can be used with other publicly available comparative analysis software tools or they can be saved as text files for use with "Batch BLAST Search".

Case study: *Haemophilus parasuis* genome data

To demonstrate the functionalities of DraGnET we used the web application to store genomic data from three strains of *Haemophilus parasuis*, two draft genomes (strains 29755 and 12939) and a complete reference genome (strain SH0165) [46], and to perform preliminary cross strain comparisons to identify protein products common to each strain. *H. parasuis* is a bacterial pathogen that causes severe respiratory disease in swine and vaccines effective against multiple isolates are lacking [47]. Since outer membrane proteins, including lipoproteins, that are shared among the *H. parasuis* strains represent potential broadly protective antigens, identifying common genes is a first step toward vaccine development. Draft genome sequence data for strains 29755 and 12939 were generated using the Illumina/Solexa Genome Analyzer II platform (G. Phillips, D. Dyer, and K. Register, unpublished data). The genomes were assembled using SH0165 as a reference genome using NextGene software (State College, PA). Annotation was performed through the Institute for Genome Sciences (IGS) Annotation Engine offered by the University of Maryland, School of Medicine.

Initially, annotated genome sequence data representing the three strains were formatted for use with DraGnET by conversion to semicolon-separated files. Subsequently, information for each strain was entered and the corresponding file was uploaded and populated in the database using the application's web interface (Figure 4). Once in the database, the annotated data was available for searching and modifying. As shown in Figure 7, "Quick Search" was used to search for information related to the gene identifier "hph_875"; which returned a table with the annotated gene information. Data modification is an important functionality provided through DraGnET, especially in the case of draft genome data. To demonstrate this capability, gene information related to gene identifier "hph_1391" was selected for updating. As shown in Figure 6, the following gene attributes were selected and modified: gene description, localization, and signal sequence. Once submitted, all modifications to the data were confirmed using "Quick Search". DraGnET provides additional functionalities for preliminary analysis of draft and complete genome data. To identify protein products common to all three of the *H. parasuis* strains, "Advanced Search" and BLAST functionalities provided through the DraGnET interface were used to perform preliminary cross strain comparisons. This demonstrates the DraGnET application is ideally suited for smaller companies or academic labs that are just beginning to use next-generation sequencing for vaccine development.

DISCUSSION

While data from genome sequencing projects typically become publicly available through sequence repositories, the rate at which large-scale sequence information is being generated and subsequent analysis will, in many cases, delay public availability of the data. In addition, sequencing projects where proprietary data are generated are limited as to how the information

can be managed and analyzed until it is ready for public reposition. This limitation emphasizes the need for software applications that provide researchers with in-house data management and analysis capabilities. While some of the features of DraGnET are provided with other applications, our software provides a user friendly in-house web application that enables researchers to manage their own unpublished or proprietary annotated draft genome data at the initial stage of development without having prior knowledge of query languages necessary for data storage, retrieval and update.

Additional features of the application include BLAST capabilities and the automatic generation of FASTA files from protein sequence data stored in the database. A web interface is provided for use of stand-alone BLAST alleviating the requirement to perform searches through command line and allowing users to search against a single strain or multiple strains as well as perform cross strain comparisons once the BLAST database files are created. DraGnET was designed to store and compare different strains from the same species; however the web interface design is generic enough to accommodate multiple organisms and their related strains. Additionally, the DraGnET software can be further developed to customize the program for specific needs.

As demonstrated in the case study, DraGnET provides researchers with an application that can be used as a first step toward data curation and analysis. Subsequently, after the data are made publicly available, more comprehensive analysis may be performed, for example by any of the aforementioned analysis software. Alternatively, the sequence data can continue to be analyzed using in-house programs, including annotation and BLAST comparisons [37, 48].

Currently gene attributes selected for storage and use with DraGnET are fixed. Further development of DraGnET will include the storage of more comprehensive annotation data as

well as more advanced functionalities for comparative analysis.

DraGnET currently contains draft and complete genome data from three strains of *H. parasuis* made available for collaborative research efforts. Readers are encouraged to visit the DraGnET website located at http://www.dragnet.cvm.iastate.edu and examine the functionalities of the software.

CONCLUSIONS

New genome sequencing methods now allow multiple draft genomes to be generated, assembled, and annotated at an unprecedented rate at modest expense. Following sequencing, assembly and annotation, there is an immediate need for the data to be organized, stored, curated and formatted for comparative analysis. The DraGnET software is an ideal in-house tool that allows i.) storage and integration of annotated data generated from different annotation platforms in a database, ii.) retrieval of gene and strain information based upon basic or advanced search parameters, iii.) management of gene and strain information, iv.) generation of FASTA formatted files for all strains stored in the database, v.) sequence similarity searches using BLAST, vi.) Batch BLAST searches for cross strain comparisons and vii.) retrieval of strain specific genes based upon Batch BLAST results. The application allows for the setup of individual projects used on local machines or may be deployed through Internet (or Intranet) access for use by other researchers across different locations. To demonstrate this, we setup a DraGnET project, deployed it for Internet access, and identified potential vaccine targets in multiple strains of *H. parasuis* using preliminary cross strain comparisons.

AVAILABILITY AND REQUIREMENTS

- Project name: Draft Genome Evaluation Tool (DraGnET)
- Project home page: http://www.dragnet.cvm.iastate.edu
- Operating system(s): Microsoft Windows 2003 and Windows XP
- Programming language: Java
- Other requirements: JRE 1.6.0, MySQL 5.0, MySQL 5.0GUI Tools, Apache Tomcat 6.0 and Blast 2.2.18
- License: GNU GPL

ACKNOWLEDGEMENTS

The authors thank Fadi Towfic for helpful suggestions and Josh Mack for technical support. The authors also thank Dr. Michelle Giglio for support during the annotation process performed through the Institute for Genome Sciences (IGS) Annotation Engine offered by the University of Maryland, School of Medicine http://ae.igs.umaryland.edu/cgi/index.cgi. This work was funded in part by the Iowa Healthy Lifestock Initiative and the National Pork Board.

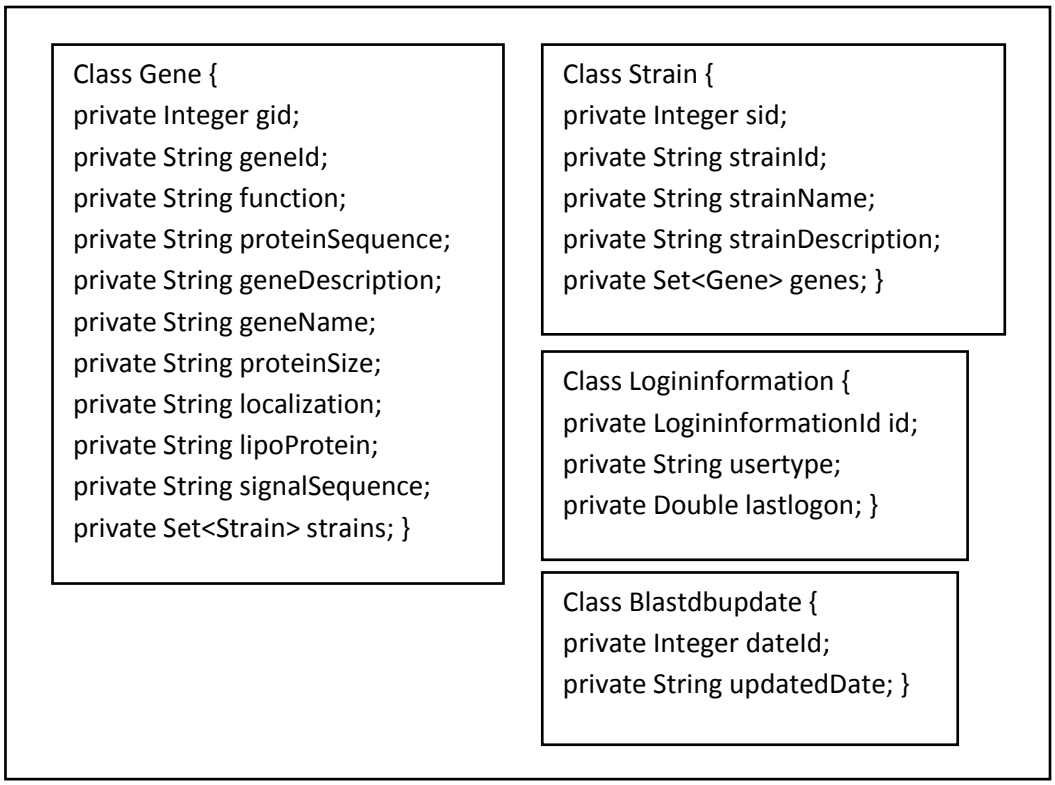

Figure 1: Java classes
Four Java classes are used to define the Gene, Strain, Logininformation and Blastdbupdate objects. The Gene class defines variables for the following annotated gene information: gene identification (gid and geneId), gene function (function), the protein sequence (proteinSequence), a description of the gene (geneDescription), the name of the gene (geneName), the size of the protein (proteinSize), the subcellular localization (localization), if the protein is predicted to be a lipoprotein (lipoprotein), if the protein is predicted to have a signal sequence (signalSequence) and the set of strains that contain the genes (Set<Strain> strains). The Strain class defines variables for strain information such as a strain identifier (sid and strainId), the strain name (strainName), a description of the strain (strainDescription), and the set of genes contained in the strain (Set<Gene> genes). The Logininformation class defines variables for the user login identifier (LogininformationId id), the usertype and the time the user logged in (lastlogon). The Blastdbupdate class defines variables for the date the last update was made to the data (dateId and updatedDate).

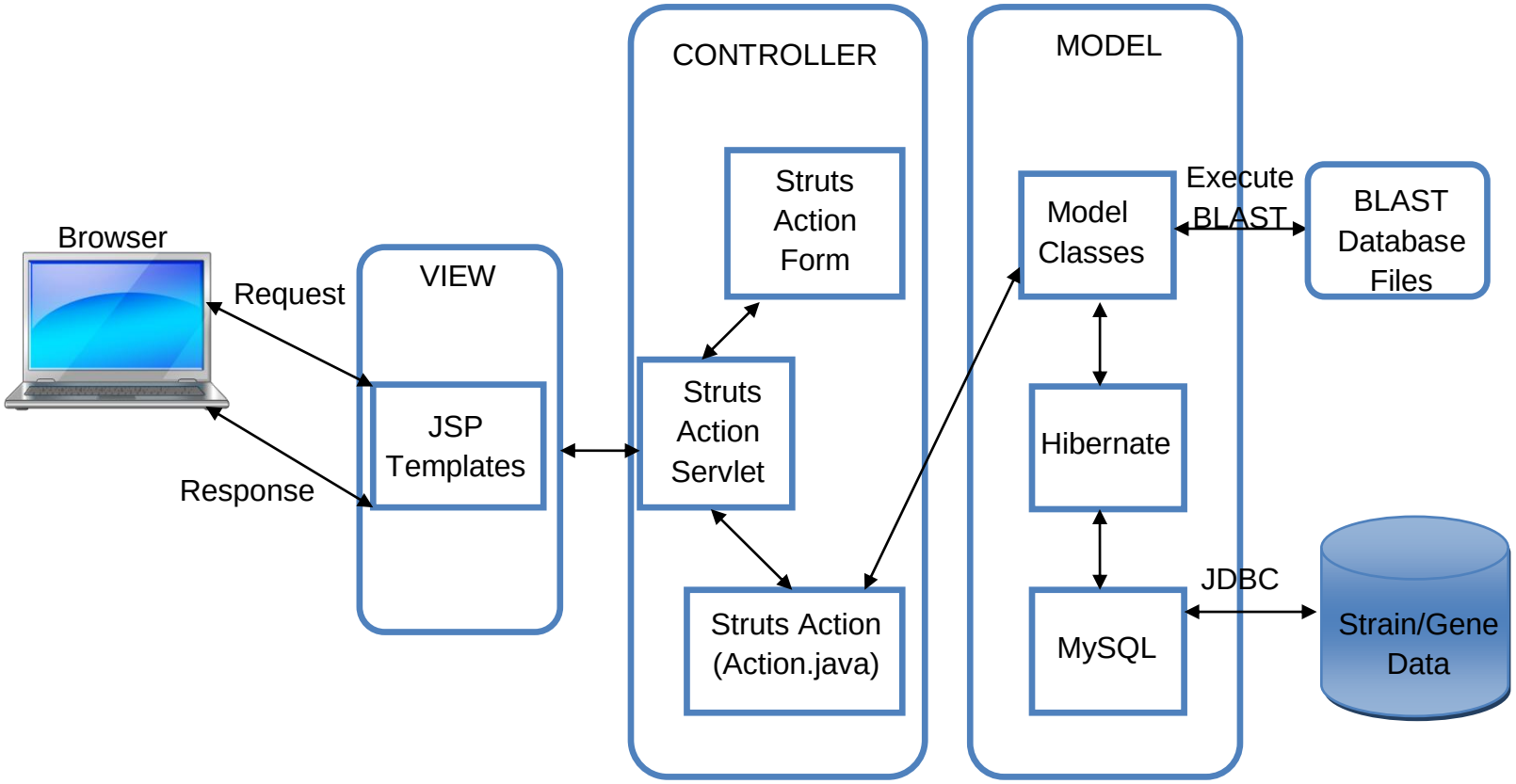
CONTROLLER
MODEL
Browser
Request
Response
VIEW
JSP
Templates
Struts
Action
Form
Struts
Action
Servlet
Struts Action
(Action.java)
Model
Classes
Execute
BLAST
BLAST
Database
Files
Hibernate
MySQL
JDBC
Strain/Gene
Data

Figure 2: DraGnET software architecture

The DraGnET web application uses Struts to implement the Model-View-Controller (MVC) architecture. The view represents the presentation of the application and is implemented through Java Server Pages (JSP). The Controller is responsible for intercepting and translating user input into actions to be performed by the Model. The Controller receives the request from the browser, invokes a business operation and coordinates the view to be returned to the browser. The Struts Action servlet populates information from the JSP to the appropriate Struts Action Form then throws control to the Struts Action. The Struts Action gets data from the appropriate Struts Action Form and sends the information to the Model where certain actions like retrievals and updates will be performed. The Model is where communication with the database takes place through Hibernate. Hibernate is used to map Model Classes (Java objects) to tables in the database. Model Classes are also used to execute BLAST functionalities provided through the application's web interface. The Model represents enterprise data and the business rules that govern access to and updates of this data.

Figure 3: Web interface- DraGnET Home Page

Listed on the DraGnET home page are links for downloading the DraGnET setup package ("DraGnET Application Setup Package"), testing search and BLAST capabilities ("Example Files"), generating FASTA formatted files ("Generate FASTA files") and all "Search" functionalities.

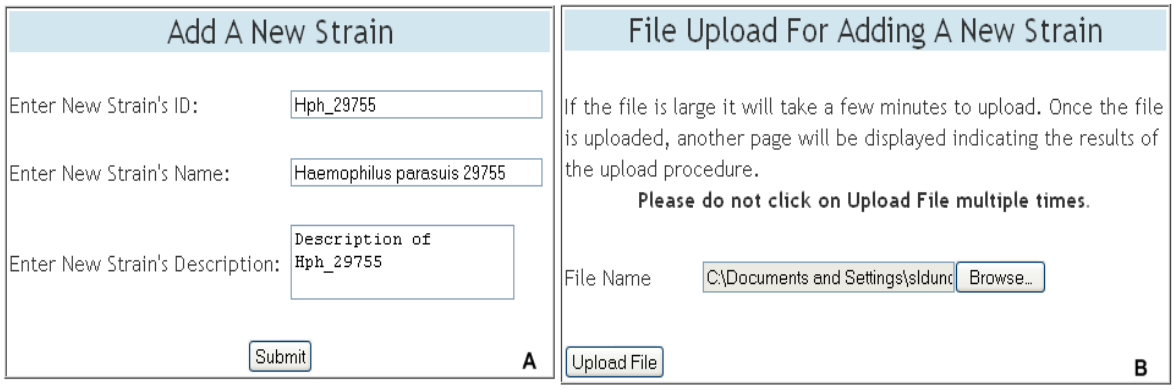

Figure 4: Adding a new strain

The data entry tables displayed on the web pages for inserting a new strain. In the first table (A) the curator enters the strain Id, strain name and strain description of the new strain. In the second table (B) the curator is directed to upload a file containing gene information for genes contained in the strain. The strain and gene information is then stored in the database.

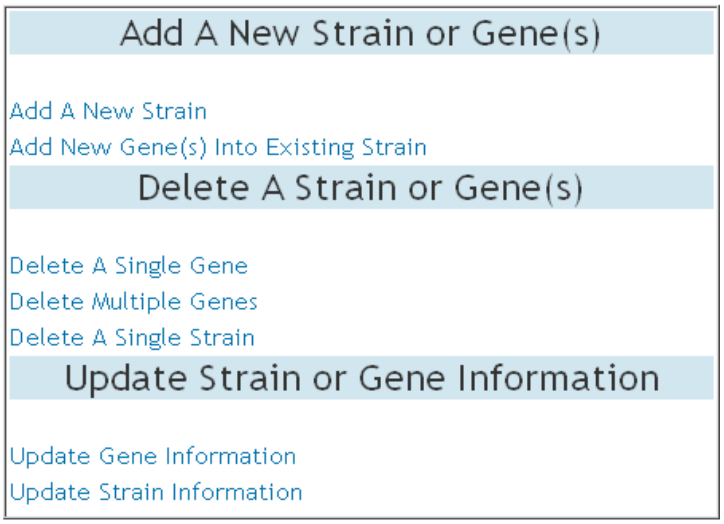

Figure 5: Data Modification

The table displayed on the web page for modifying gene and strain data. As shown in the table, modifications that can be made by the curator to gene and strain data include adding, deleting and updating gene or strain information.

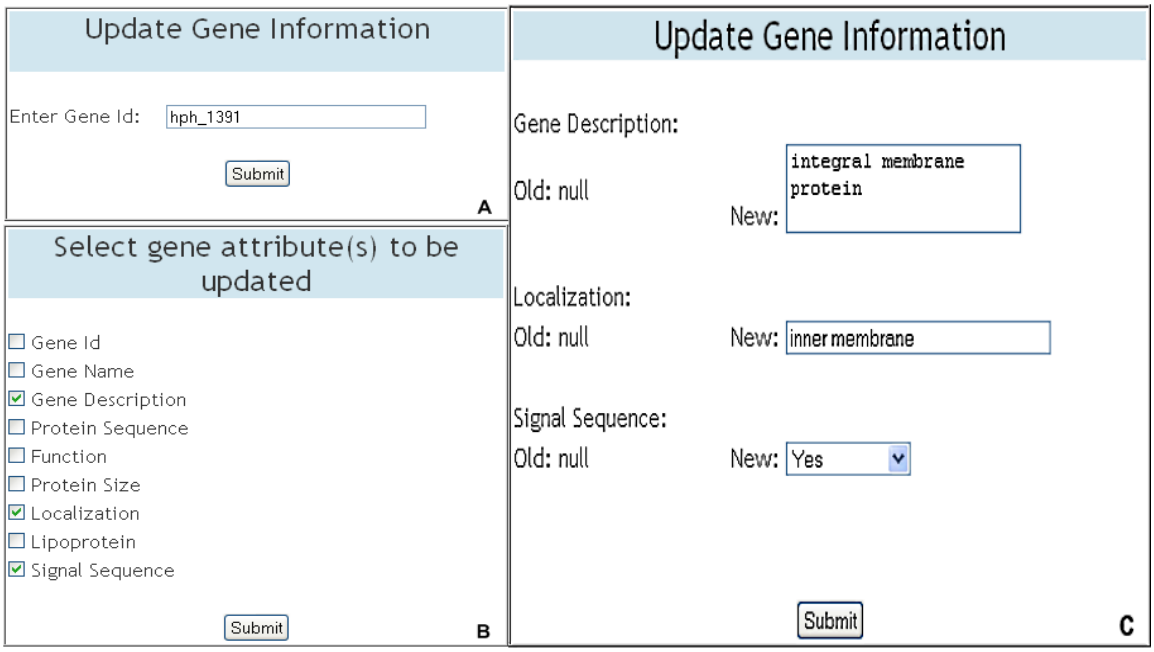

Figure 6: Updating gene information

The data entry tables displayed on the web pages for updating gene information. In the top left table (A) the gene Id of the gene whose information needs modification is entered and submitted. The table on the bottom left (B) allows the curator to select gene attributes that need to be modified. Subsequently, as shown in table (C), for each attribute selected, the gene information currently stored in the database is displayed on the left side of the table as "Old" information and on the right side of the table changes to the gene information may be entered under "New".

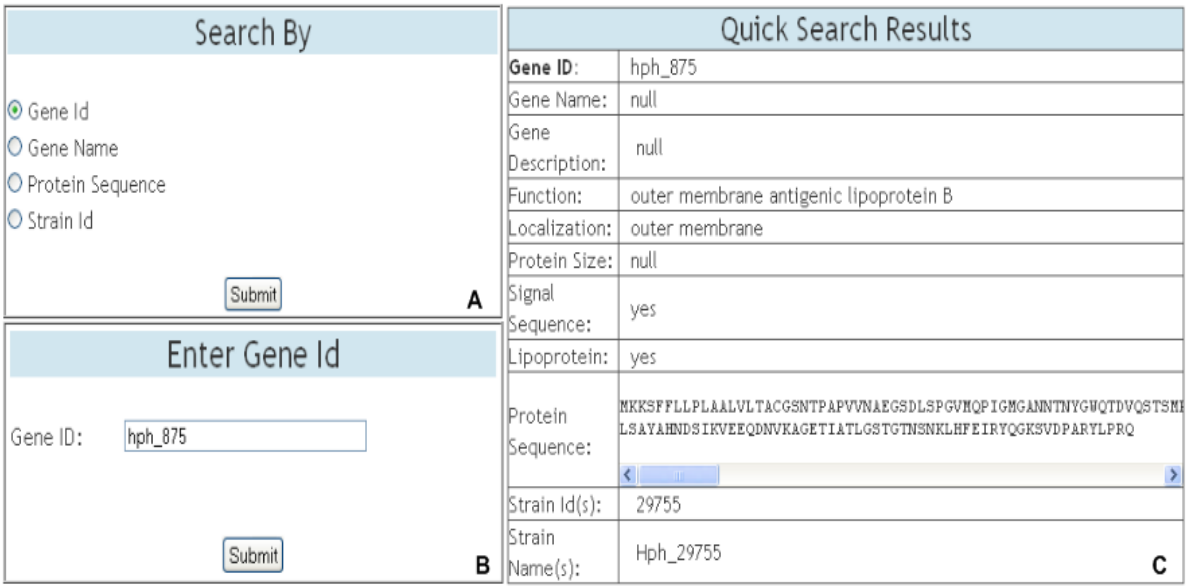

Figure 7: Quick Search
The data entry tables and results table displayed on the web pages for "Quick Search". In the top left table (A) the user selects a gene or strain search attribute. In the bottom left table (B) the user enters information for the chosen search criteria. Subsequently, the results table (C) displays information for the chosen gene or strain.

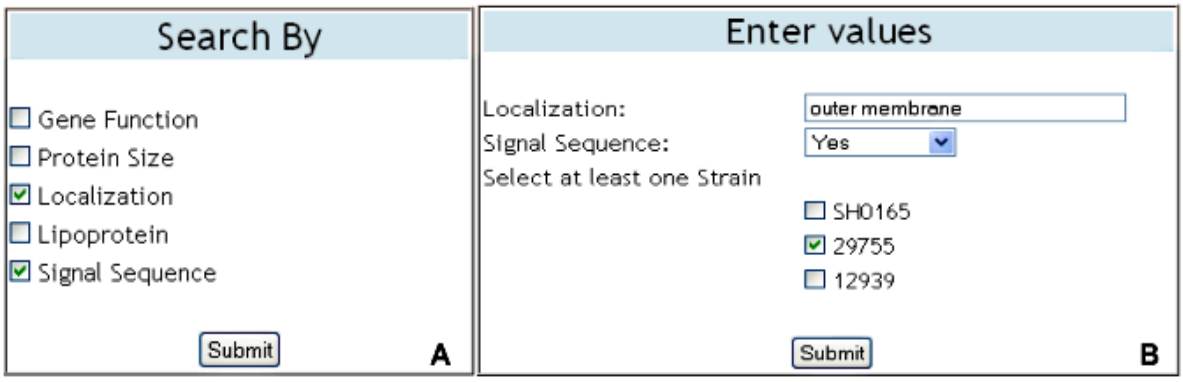

Figure 8: Advanced Search
The data entry tables displayed on the web pages for "Advanced Search". Using the table on the left (A) users may customize their search for gene information by selecting single or multiple search attributes. The table on the right (B) allows users to enter and select values for the chosen search criteria. Subsequently, a text file containing search results is available for download.

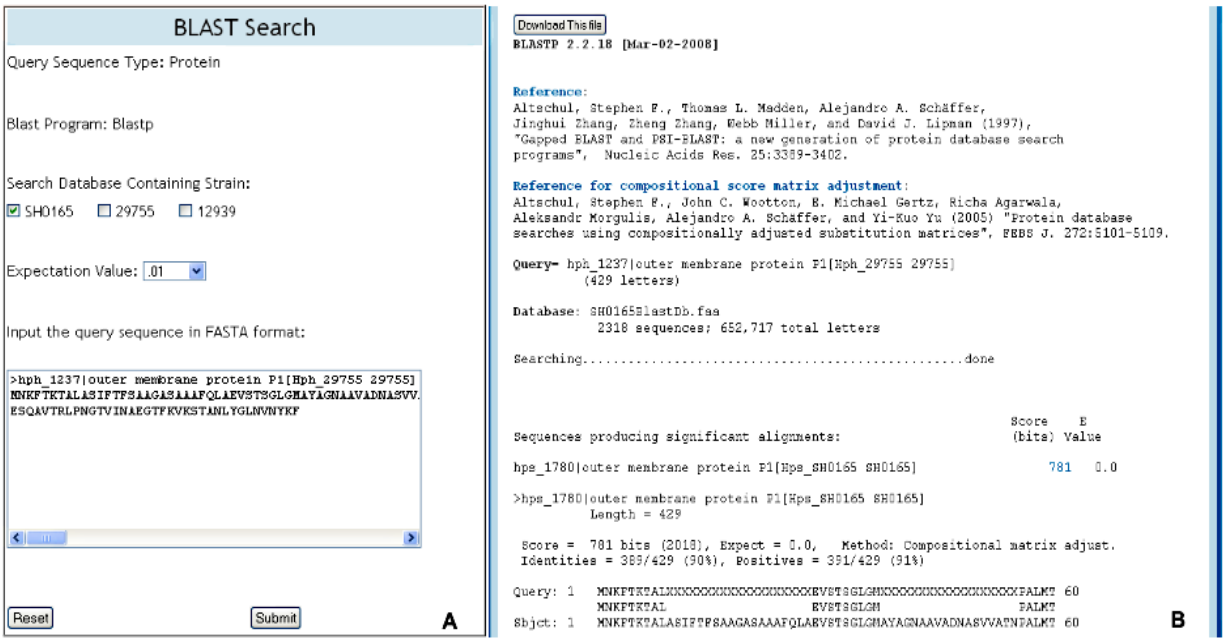

Figure 9: BLAST Search

The data entry table and results displayed on the web pages for "BLAST Search". As shown in table (A), users have the option to select a single or multiple search database(s). In this example we have three BLAST databases available for searching that represent strains SH0165, 29755 and 12939. To refine their search, users have the option to change the E-value with the default being .01. A text box is provided for users to enter a FASTA formatted protein query sequence. The results (B) are displayed to the user and are available for download.

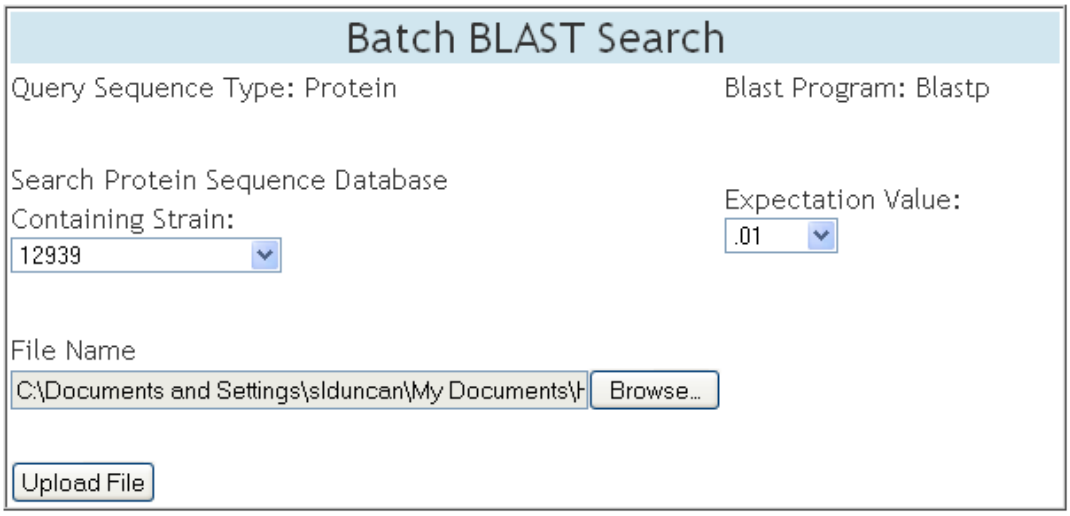

Figure 10: Batch BLAST Search
The data entry table displayed on the web page for "Batch BLAST Search". Users select a single search database from a drop down menu. In this example, the BLAST database representing strain 29755 was chosen. To refine their search, users have the option to change the E-value with the default being .01. Users then upload a text file containing FASTA formatted protein sequences which will be used as the set of query sequences. The results format is the same as "BLAST Search".

REFERENCES

1. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J Mol Biol* 1975, **94:**441-448.
2. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc Natl Acad Sci U S A* 1977, **74:**5463-5467.
3. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
4. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G: **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies.** *Nucleic Acids Res* 2006, **34:**e22.
5. Turcatti G, Romieu A, Fedurco M, Tairi AP: **A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.** *Nucleic Acids Res* 2008, **36:**e25.
6. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science* 2005, **309:**1728-1732.
7. Mardis ER: **Next-generation DNA sequencing methods.** *Annu Rev Genomics Hum Genet* 2008, **9:**387-402.
8. Marguerat S, Wilhelm BT, Bahler J: **Next-generation sequencing: applications beyond genomes.** *Biochem Soc Trans* 2008, **36:**1091-1096.
9. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5:**16-18.
10. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.
11. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome.** *Curr Opin Genet Dev* 2005, **15:**589-594.
12. Field D, Wilson G, van der Gast C: **How do we compare hundreds of bacterial genomes?** *Curr Opin Microbiol* 2006, **9:**499-504.
13. Fukiya S, Mizoguchi H, Tobe T, Mori H: **Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray.** *J Bacteriol* 2004, **186:**3911-3921.
14. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102:**13950-13955.
15. Bjorkholm B, Lundin A, Sillen A, Guillemin K, Salama N, Rubio C, Gordon JI, Falk P, Engstrand L: **Comparison of genetic divergence and fitness between two subclones of Helicobacter pylori.** *Infect Immun* 2001, **69:**7832-7838.
16. Fitzgerald JR, Sturdevant DE, Mackie SM, Gill SR, Musser JM: **Evolutionary genomics of Staphylococcus aureus: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic.** *Proc Natl Acad Sci U S A* 2001, **98:**8821-8826.
17. Muzzi A, Masignani V, Rappuoli R: **The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials.** *Drug Discov Today* 2007, **12:**429-439.

18. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312:**1355-1359.
19. Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C: **Archaea predominate among ammonia-oxidizing prokaryotes in soils.** *Nature* 2006, **442:**806-809.
20. Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F: **Metagenomic analysis of the microbial community associated with the coral Porites astreoides.** *Environ Microbiol* 2007, **9:**2707-2719.
21. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V, et al: **Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach.** *BMC Genomics* 2006, **7:**246.
22. Cheung F, Haas BJ, Goldberg SM, May GD, Xiao Y, Town CD: **Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology.** *BMC Genomics* 2006, **7:**272.
23. Torres TT, Metta M, Ottenwalder B, Schlotterer C: **Gene expression profiling by massively parallel sequencing.** *Genome Res* 2008, **18:**172-177.
24. Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB: **Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing.** *Plant Physiol* 2007, **144:**32-42.
25. Medigue C, Moszer I: **Annotation, comparison and databases for hundreds of bacterial genomes.** *Res Microbiol* 2007, **158:**724-736.
26. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, et al: **The integrated microbial genomes (IMG) system.** *Nucleic Acids Res* 2006, **34:**D344-348.
27. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM, Dubchak I, Anderson I, Lykidis A, Mavromatis K, et al: **The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions.** *Nucleic Acids Res* 2008, **36:**D528-533.
28. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC: **IMG ER: a system for microbial genome annotation expert review and curation.** *Bioinformatics* 2009, **25:**2271-2278.
29. Romualdi A, Felder M, Rose D, Gausmann U, Schilhabel M, Glockner G, Platzer M, Suhnel J: **GenColors: annotation and comparative genomics of prokaryotes made easy.** *Methods Mol Biol* 2007, **395:**75-96.
30. Romualdi A, Siddiqui R, Glockner G, Lehmann R, Suhnel J: **GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness.** *Bioinformatics* 2005, **21:**3669-3671.
31. Uchiyama I: **MBGD: microbial genome database for comparative analysis.** *Nucleic Acids Res* 2003, **31:**58-62.
32. Uchiyama I: **MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups.** *Nucleic Acids Res* 2007, **35:**D343-346.
33. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The Comprehensive Microbial Resource.** *Nucleic Acids Res* 2001, **29:**123-125.

34. Blom J, Albaum SP, Doppmeier D, Puhler A, Vorholter FJ, Zakrzewski M, Goesmann A: **EDGAR: a software framework for the comparative analysis of prokaryotic genomes.** *BMC Bioinformatics* 2009, **10:**154.
35. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Medigue C: **MaGe: a microbial genome annotation system supported by synteny results.** *Nucleic Acids Res* 2006, **34:**53-65.
36. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al: **The RAST Server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9:**75.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
38. Rappuoli R: **Reverse vaccinology, a genome-based approach to vaccine development.** *Vaccine* 2001, **19:**2688-2691.
39. Serruto D, Serino L, Masignani V, Pizza M: **Genome-based approaches to develop vaccines against bacterial pathogens.** *Vaccine* 2009, **27:**3245-3250.
40. Bambini S, Rappuoli R: **The use of genomics in microbial vaccine development.** *Drug Discov Today* 2009, **14:**252-260.
41. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R, et al: **Identification of a universal Group B streptococcus vaccine by multiple genome screen.** *Science* 2005, **309:**148-150.
42. Myers GS, Parker D, Al-Hasani K, Kennan RM, Seemann T, Ren Q, Badger JH, Selengut JD, Deboy RT, Tettelin H, et al: **Genome sequence and identification of candidate vaccine antigens from the animal pathogen Dichelobacter nodosus.** *Nat Biotechnol* 2007, **25:**569-575.
43. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, et al: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing.** *Science* 2000, **287:**1816-1820.
44. Al-Hasani K, Boyce J, McCarl VP, Bottomley S, Wilkie I, Adler B: **Identification of novel immunogens in Pasteurella multocida.** *Microb Cell Fact* 2007, **6:**3.
45. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
46. Yue M, Yang F, Yang J, Bei W, Cai X, Chen L, Dong J, Zhou R, Jin M, Jin Q, Chen H: **Complete genome sequence of Haemophilus parasuis SH0165.** *J Bacteriol* 2009, **191:**1359-1360.
47. Rapp-Gabrielson VJ, S. R. Oliveira, and C. Pijoan: **Haemophilus parasuis.** In *Diseases of Swine*. 8th edition. Edited by B. E. Straw JJZ, S. D'Allaire, and D. J. Taylor. Ames, I.A.: Blackwell Publishing; 2006: 475-481
48. Berriman M, Rutherford K: **Viewing and annotating sequence data with Artemis.** *Brief Bioinform* 2003, **4:**124-132.

CHAPTER 6. General Conclusions

Structural determinations of the SRP and receptor have provided details of how components of this essential and highly conserved protein localization machinery interact in precise detail (1, 2, 6, 7, 9, 10, 14). In addition, these efforts have led to several hypotheses as to how the components function in protein targeting. It has been the goal of the studies described in this dissertation to use the structural information to guide experiments designed to better understand how the structure of Ffh directs the function of the SRP. For this, we have taken multiple genetic approaches that, along with bioinformatic analyses, have led to new insights into how specific domains of Ffh contribute to its activity *in vivo*.

Specifically, we have developed a genetic system that allowed complementation tests to be performed using an *E. coli* strain that conditionally grows at 42^{o}C only if it is transformed with a plasmid expressing a functional copy of *ffh*. This strain has allowed us to isolate and test mutants in both the finger loop domain and the M-domain of Ffh to understand the importance of specific amino acids in each region of the protein. In addition, we have overcome a limitation with performing complementation test using multiple copy number plasmids by utilizing a system to express mutant alleles in single copy so that their expression will be more physiologically relevant (13).

Due to the flexibility and hydrophobic nature of the finger loop domain of Ffh, structural information is lacking and has therefore been difficult to study. However, by screening a random sequence library (4, 11, 12, 15), we were able to select sequences that

could replace the wild type finger loop domain and provide a functional SRP. Bioinformatic analyses of several of these mutants revealed two important features, including the necessity of maintaining a highly hydrophobic N-terminus along with a decrease in hydrophobicity toward the C-terminal end. Comparison of the finger loop domain from multiple species also revealed that while there is considerable variation in the residue composition of the finger loop toward the C-terminus, the physicochemical properties of the amino acids are similar. Our results support the proposal that the finger loop domain contributes to the hydrophobic interactions in signal sequence binding and provides additional flexibility to the signal sequence binding domain of Ffh required for accommodating a variety of signal sequences (3, 5, 8). The observed periodic distribution of hydrophobic amino acids and the gradient of hydrophobicity throughout the finger loop domain are in support of the contribution of flexibility this domain provides to Ffh signal sequence binding.

To further characterize the M-domain of Ffh, we turned our focus to the importance of methionine residues. Although the importance of "methionine bristles" had been suggested for Ffh function, there have been no direct tests of this hypothesis. It had been previously observed that methionine residues found in the M-domain of Ffh from *E. coli* were substituted in *T. aquaticus* with less flexible hydrophobic amino acids such as leucine, isoleucine and phenylalanine (9). By performing a more extensive phylogenetic sequence analysis of Ffh from bacteria and archaea representing varying optimal growth temperatures (OGT), including hyperthermophiles ($\geq 75^{o}$C), thermophiles (between 40^{o}C and 75^{o}C), psychrophiles (between 10^{o}C and 15^{o}C) and mesophiles (between 25^{o}C and 40^{o}C) we observed that indeed methionine residues are found at a

higher percentage in psychrophiles and mesophiles than hyperthermophiles and thermophiles.

As a direct test of the importance of methionines in Ffh function, we also constructed a series of mutants by substituting all of the methionines within the αM4 helix at the extreme C-terminus of Ffh with several other amino acids. For this, we selected additional hydrophobic residues, including alanine, cysteine, leucine, isoleucine, phenylalanine, valine, tyrosine and tryptophan, as well the charged amino acid glutamate. By expressing mutant *ffh* alleles in single and multiple copy, we were able to clearly determine that residues valine, phenylalanine, tyrosine, tryptophan, and isoleucine could function to replace methionine residues found in αM4 helix and the extreme C-terminal region of the M-domain. Surprisingly, we observed that the *ffh*α4M→V mutant grew near wild type levels, while substitutions of leucine and isoleucine, both having side chains more similar to methionine, were extremely poor replacements. Mutant *ffh*α4M→V was able to support cell viability at several growth temperatures (30^{o}C, 37^{o}C and 42^{o}C) when present in single and multiple copy; however, *ffh*α4M→I, when present in single copy, only supported growth of SLD106 at 42^{o}C. Additionally, elevated expression of the product of *ffh*α4M→L was required to support growth of SLD106 at 42^{o}C, albeit the colonies were unable to form single colonies upon restreaking.

From these unexpected results, we tested the ability of valine to replace all of the methionines in the complete M-domain. While this mutant failed to support growth, addition of 3 highly conserved methionines (positions 344, 383 and 426 corresponding to the *E. coli* Ffh protein) restored SRP function. Further studies will be required to determine if any of these methionine residues are necessary for Ffh function.

In conclusion, our results provide further insights into characteristics of the signal sequence binding domain of Ffh and call into question the necessity of methionine residues found in the M-domain of Ffh from *E. coli*. Future research directions include additional mutational analysis of three highly conserved methionine residues found in the M-domain in order to determine the importance of these residues in signal sequence recognition. In addition, biochemical analysis of the finger loop mutants will be informative in order to identify how changes in this unstructured domain affect SRP function and to further elucidate its role in Ffh function.

REFERENCES

1. **Batey, R. T., R. P. Rambo, L. Lucast, B. Rha, and J. A. Doudna.** 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. Science **287:**1232-9.
2. **Clemons, W. M., Jr., K. Gowda, S. D. Black, C. Zwieb, and V. Ramakrishnan.** 1999. Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 Å resolution: evidence for the mechanism of signal peptide binding. J. Mol. Biol. **292:**697-705.
3. **Driessen, A. J., and N. Nouwen.** 2008. Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. **77:**643-67.
4. **Dube, D. K., M. E. Black, K. M. Munir, and L. A. Loeb.** 1993. Selection of new biologically active molecules from random nucleotide sequences. Gene **137:**41-7.
5. **Egea, P. F., R. M. Stroud, and P. Walter.** 2005. Targeting proteins to membranes: structure of the signal recognition particle. Curr Opin Struct Biol **15:**213-20.
6. **Ilangovan, U., S. H. Bhuiyan, C. S. Hinck, J. T. Hoyle, O. N. Pakhomova, C. Zwieb, and A. P. Hinck.** 2008. *A. fulgidus* SRP54 M-domain. J. Biomol. NMR **41:**241-8.
7. **Janda, C. Y., J. Li, C. Oubridge, H. Hernandez, C. V. Robinson, and K. Nagai.** 2010. Recognition of a signal peptide by the signal recognition particle. Nature.
8. **Keenan, R. J., D. M. Freymann, R. M. Stroud, and P. Walter.** 2001. The signal recognition particle. Annu. Rev. Biochem. **70:**755-75.

9. **Keenan, R. J., D. M. Freymann, P. Walter, and R. M. Stroud.** 1998. Crystal structure of the signal sequence binding subunit of the signal recognition particle. Cell **94:**181-91.
10. **Oh, D. B., G. S. Yi, S. W. Chi, and H. Kim.** 1996. Structure of a methionine-rich segment of *Escherichia coli* Ffh protein. FEBS Lett. **395:**160-4.
11. **Palzkill, T., Q. Q. Le, A. Wong, and D. Botstein.** 1994. Selection of functional signal peptide cleavage sites from a library of random sequences. J. Bacteriol. **176:**563-8.
12. **Patel, P. H., and L. A. Loeb.** 2000. DNA polymerase active site is highly mutable: evolutionary consequences. Proc. Natl. Acad. Sci. USA **97:**5095-100.
13. **Peterson, J. M., and G. J. Phillips.** 2008. Characterization of conserved bases in 4.5S RNA of *Escherichia coli* by construction of new F' factors. J. Bacteriol. **190:**7709-18.
14. **Rosendal, K. R., K. Wild, G. Montoya, and I. Sinning.** 2003. Crystal structure of the complete core of archaeal signal recognition particle and implications for interdomain communication. Proc. Natl. Acad. Sci. USA **100:**14701-6.
15. **Skandalis, A., and L. A. Loeb.** 2001. Enzymatic properties of rat DNA polymerase beta mutants obtained by randomized mutagenesis. Nucleic Acids Res. **29:**2418-26.